Within Genome Variation of Germ-line and Somatic Mutation

Thomas C. A. Smith

Ph.D. Biology

University of Sussex

September 2016

Table of Contents

Declaration

I hearby declare that this thesis has not been, and will not be, submitted in whole or in part to any other university for the award of an other degree.

Signature.....

Acknowledgements

I would like to thank my supervisor, Adam Eyre-Walker, for giving me the opportunity to explore this exciting topic for his help and support over the past 4 years. I am especially grateful to him for patiently eduring some of my more whimsical ideas, and for fostering a culture of freedom in the group. This has allowed me to develop interests I would otherwise have not encountered and allowed me to focus on developing an array of new and exciting skills. His support has been instrumental in my success over the course of my Ph.D.

I would also like to thank my co-supervisor, Antony Carr, for his continual encouragement, who along with Yasu has provided interesting conversations relating to cancer biology, among other topics. These conversations have provided the opportunity for me to approach my work from a different perspective, which has no doubt enabled me to have a more holistic understanding of the themes covered in this thesis.

I would also like to thank my office mates, David, Keith, Jenny, Alex, Dan and Beth, for providing a vibrant, and generally studious, environment in which to work - the addition of a guitar as part of the office furniture over the last 6 months has been a surprisingly welcome addition, and along with my many sporting friends, has been essential in providing much needed relief and distraction at times of stress.

Finally, I would like to thank my family. My mother, my father and my sister and her children for providing unquestioning support and open, welcoming homes when required. My brother for the evenings spent in pubs and providing a welcome distraction. And of course, Justyna, for her support, enouragement and understanding; providing the space for me to focus on my work without distraction, something that has been especially important in the latter months of writing.

Preface

The research presented here was carried out at the University of Sussex. Author contribution and publication details are as follows:

Chapter 2 has been published as: Smith. T., *et al.* 2016. Extensive Variation in the Mutation Rate Between and Within Human Genes Associated with Mendelian Disease. Human Mutation Vol. 37, No 5:488–494. This study was desgined by Adam Eyre-Walker (AEW). AEW conducted the parameter estimation and I conducted all other analyses. AEW and I wrote the final manuscript.

Chapter 3 has been accepted for publication to PeerJ as: Smith. T., Carr. A. M. and Eyre-Walker. A., 2016. Are sites with multiple single nucleotide variants in cancer genomes a consequence of drivers, hypermutable sites or sequencing errors? This is currently in publication. AEW and I designed the study with input for Antony Carr (AMC). AEW carried out the parameter estimation. and I conducted all other analyses. AEW and I wrote the final manuscript with input for AMC.

For chapter 4, I designed the study with input from AMC and AEW. I conducted all the analyses and wrote the chapter with input from AMC and AEW.

Chapter 5 was designed by AEW and I. I collected and processed all data. AEW calculated the expected correlations and parameter estimations. I conducted all other analyses. AEW and I wrote the final manuscript.

Appendix 7 was designed by me with input from AEW and AMC. I performed alll analyses and wrote the final muanuscript with input from AEW and AMC.

Summary

Variation in the mutation rate along the human genome, if not properly understood and accounted for, has the potential to confound evolutionary studies, produce spurious driver candidates in cancer studies, and hinder the diagnostics aimed at understanding the etiologies of genetic diseases.

In this thesis I consider mutation rate variation in both the germ-line and somatic tissues, at varying scales. In the germ-line, the major advance of this thesis over previous works is use of direct methods to analyse the magnitude, scale and determinants of mutation rate variation. This has enabled us to tease apart the evolutionary and mutational forces, whilst directly quantifying the variation in the human mutation rate at different scales; at large scales the variation appears to be quite modest, however at the single nucleotide scale there is potentially huge cryptic variation in the mutation rate. I envisage that within the near future, the increase in *de novo* mutations coming from pedigree studies will allow for even greater understanding.

I extend this work into somatic tissues, however due to the quality of data and heterogeneity of samples and cell types, the primary findings lean towards highlighting areas of improvement for next genereation sequencing (NGS) pipelines - I show ~4% of all single nucleotide variants from cancers appear to be errors - and develop methods with which future studies could provide more insight. With these methods and the ever increasing flow of somatic single nucleotide variants, coupled with the continual improvements in NGS technology, it should soon be possible to provide accurate answers to the questions posed of somatic mutation rate variation in this thesis.

1. Introduction.

1.1 General introduction

Mutation is an important phenomenon in biology, without it genetic variation can not arise in a population and evolution can not proceed. Mutation is also the source of genetic disease. In *Homo sapiens* (humans) deleterious germ-line mutations cause monogenic diseases and influence susceptibility to polygenic diseases such as autism and schizophrenia (Awadalla *et al.* 2010; Conrad *et al.* 2011; Iossifov *et al.* 2012; Iossifov *et al.* 2014; Neale *et al.* 2012; O'Roak *et al.* 2012; O'Roak *et al.* 2011; Sanders *et al.* 2012; Veltman & Brunner 2012). It follows that the rate at which mutations occur in an individual, and any variation in that rate, will impact both the evolution of that individual's species and the prevalence of disease in both the individual and their species at a population level. In somatic tissues deleterious mutations are responsible for diseases such as cancer along with potentially contributing to a myriad of neurological disorders (Poduri *et al.* 2012; Poduri *et al.* 2013; Rivière *et al.* 2012; Lee *et al.* 2012) and thus understanding the rate and spectrum of somatic mutation also has clinical importance.

The focus is upon point mutations, which are the most abundant form of mutations in both the the germ-line and soma. Nucleotides are classified by their structure as either pyrimidines; cytosine (C) and thymine (T) or purines; adenine (A) and guanine (G). Mutations occurring between the same type are known as transitions, e,g C<>T or G<>A, all other point mutations are called transversions. Transitions occur at a higher rate than transversions (Bulmer 1986; Cooper & Krawczak 1990; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000).

This introduction reviews the current understanding of human mutation rate variation, first of all in the germ-line, then in somatic tissues. This variation will be explored at different scales as it is evident that different processes are operative at different resolutions. The smallest scale at which variation can be studied involves at the single nucleotide level where the influence from neighboring nucleotides is very strong (Bulmer 1986; Cooper & Krawczak 1990; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000). The next scale I consider is small scale regional variation - between 2-2000 nucleotides - which appears to be be largely influenced by local GC content (Tomso & Bell 2003; Cohen *et al.* 2011) and chromatin dynamics (Ying *et al.* 2010). Above these scales we start to examine multiple kilo-bases (kb) and mega-bases (Mb); these are the scales that we associate with the molecular unit of heredity, the genes. At these large scales mutation rate appears to be related to a number of different features which will be discussed in section 1.3. Finally, I briefly explore variation at the largest scale, between chromosomes. As these form physically separate entities during cell division and can be differentially inherited, it is an obvious unit of scale to explore with very different reasons for exhibiting variation than the smaller scales. This introduction will also attempt to highlight the commonalities and differences between patterns of somatic and germ-line mutation, with an additional focus upon the disconnect between the inferred patterns of substitutions from comparative genomics and direct findings from de novo mutations (DNMs) from recent trio studies. I close by exploring the importance of considering scale in drawing conclusions on mutational processes.

1.2 The evolution of methods.

In 1947, through studying he emergence of the haemophilia phenotype, excessive bleeding, in a population, Haldane first measured the rates of male and female mutation at the X-linked

Haemophilia locus (Haldane 1947). This method relied upon assumptions of full penetrance and a consistent expressivity of one mutant disease causing locus. These assumptions could not be confirmed at the time and still to this day can hinder efforts that use phenotypic incidence to predict genotype. This is exemplified by the fact that we now know Haemophilia can be caused by mutations at multiple X-lined and autosomal loci, and can exhibit variable expressivity depending upon the concentration of the different coagulation factors produced by these loci (Mannuci & Tuddenheim, 2001). Since Haldanes work, technological advances have continually transformed the methods by which human germ-line mutation rates can be ascertained. However one thing that has remained relatively constant is the predominant reliance upon comparative studies from which mutation rates have been indirectly inferred. The initial comparison of cytochrome c (Margoliash 1963) and Haemoglobin (Zuckerkandl & Pauling 1965) protein sequences from different species, gave way to comparisons of small DNA sequences, such as pseduogenes (Nachman & Crowell 2000; Subramanian & Kumar 2003; Aguilar *et al.* 2005), and with the further development of shotgun sequencing, were followed by comparisons of whole genome sequences (Varki & Altheide 2005). The great power afforded us from comparative genomics is however, not without its issues. The substitution patterns upon which these inferences are based, are the result of millions of years of accumulated DNMs, selection, drift and other potential mechanisms, such as recombination and the associated bias gene conversion (BGC), a process whereby mismatches formed during recombination are repaired in favour of G or C over A or T bases (Duret & Arndt 2008). All of these can be hard to disentangle from each other. In addition the methods often require various assumptions about the neutrality of sequences or constancy of a molecular clock and generation times. Thus it is unclear exactly how well patterns of mutation derived from these indirect methods reflect the true pattern of germ-line DNMs. A more desirable method to study the patterns of germ-line mutations involves direct measurements taken from pedigrees, but until recently these studies were restricted to small regions, commonly the mitochondrial control region (Sigurgardottir *et al.*

2000) and as such were limited in power. However, with the relatively recent emergence of next generation sequencing (NGS) as an affordable technology, the scope of pedigree studies has broadened to allow direct measurements of DNMs from hundreds of parent-offspring trios/quartets (Awadalla *et al.* 2010; Conrad *et al.* 2011; O'Roak *et al.* 2011; O'Roak *et al.* 2012; Iossifov *et al.* 2012; Iossifov *et al.* 2014; Neale *et al.* 2012; Sanders *et al.* 2012; Veltman & Brunner 2012; Kong *et al.* 2012; Wong *et al.* 2016; Michaelson *et al.* 2012). This same technology has also facilitated an explosion in the sequencing of thousands of cancer genomes, resulting in more somatic mutations listed in public repositories - >40 million in the International Cancer Genome Consortium (ICGC) portal (https://icgc.org/) at the time of writing - than there are substitutions between the *Homo sapeins* and *Pan troglodytes* reference genomes. Prior to this, information about the somatic mutation rate was only available from calculations on single loci in cell lines (Araten *et al.* 2005; Herrero-Jimenez *et al.* 2000; Glaab & Tindall 1997; Lichtenauer-Kaligis *et al.* 1996; Bhattacharyya *et al.* 1995) or monogenic somatic disease prevalence (Hethcote & Knudson 1978; Fitzgerald *et al.* 1983; Shoichi Mizunol, Shaw Watanabe, Takeo Iwama 1993; Iwama 2001; Luebeck & Moolgavkar 2003; Hornsby *et al.* 2008).

1.3 Germ-line.

1.3.1 Between Genomes.

In the germ-line it has been shown that the mutation rates vary between species, for example *Mus musculus* have a rate of mutation twice that of *Homo sapiens* (Uchimura *et al.* 2015). Rates also vary between populations within a species, such as the surprising discovery that 5'-TCC-3' → 5'-TTC-3' mutations occur more frequently in European *Homo sapiens* than African populations

(Harris 2015). Mutation rates also differ between the sexes (Conrad *et al.* 2011), with the majority of point mutations arising from the male germ-line (Hurst & Ellegren 1998; Haldane 1947), an asymmetry that increases with paternal age (Neale *et al.* 2012; O'Roak *et al.* 2012; O'Roak *et al.* 2011; Iossifov *et al.* 2014; Iossifov *et al.* 2012; Michaelson *et al.* 2012; Kong *et al.* 2012; Francioli *et al.* 2015; Wong *et al.* 2016), and to a much lesser extent, maternal age (Wong *et al.* 2016). However, of primary interest here is that the mutation rate varies at different sites within a genome.

1.3.2 Comparative studies across the genome.

1.3.2.1 Whole chromosomes.

Variation occurs at different scales; At the level of whole chromosomes the greatest variation is seen between the sex chromosomes and the autosomes (Lercher *et al.* 2001). The Y chromosome mutates 1.5-2 fold faster than the autosomes (Makova & Li 2002) which in turn mutate about 1.3 fold faster than the X chromosome (McVean & Hurst 1997). This is thought to be due to the aforementioned male mutation bias and because the Y chromosome, autosomes and X chromosome respectively spend, on average, less time in the male germ-line (Lercher *et al.* 2001).

1.3.2.2 Large scale variatiom.

Comparative studies between species have ascertained that the mutation rate also varies at sub-chromosomal, but still at large scales. For example between genes (Wolfe *et al.* 1989; Matassi *et al.* 1999) or between Mb or multiple kb windows (Williams & Hurst 2000; Martin J. Lercher & Hurst

2002; Smith & Lercher 2002; M J Lercher & Hurst 2002; Hellmann *et al.* 2003; Tyekucheva *et al.* 2008; Duret & Arndt 2008; Stamatoyannopoulos *et al.* 2009; Don *et al.* 2013; Prendergast *et al.* 2007; Hodgkinson & Eyre-Walker 2011). This variation correlates with changes in GC content (Hellmann *et al.* 2005; Wolfe *et al.* 1989; Tyekucheva *et al.* 2008; Don *et al.* 2013; Chen *et al.* 2010), recombination rate (Hellmann *et al.* 2003; Hellmann *et al.* 2005; Martin J. Lercher & Hurst 2002; Tyekucheva *et al.* 2008; Duret & Arndt 2008; Don *et al.* 2013), replication time (Chen *et al.* 2010; Pink & Hurst 2010; Stamatoyannopoulos *et al.* 2009; Don *et al.* 2013), distance to telomeres (Hellmann *et al.* 2005; Tyekucheva *et al.* 2008; Chen *et al.* 2010; Don *et al.* 2013), chromatin state (Prendergast *et al.* 2007; Don *et al.* 2013) and nuclear lamina binding sites (Don *et al.*, 2013; Ananda, Chiaromonte & Makova, 2011). However the magnitude of this large scale variation is low, only differing about 2-fold between the highest and lowest mutating windows (Chimpanzee Sequencing Consortium, 2005), All features combined only explain about 50% of the overall variance, with each factor only explaining a small proportion (Chen *et al.* 2010; Hodgkinson & Eyre-Walker 2011; Tyekucheva *et al.* 2008). The use of Mb windows has become something of a standard for genomic analyses. It has been argued that in the germ-line this is the most pertinent scale to study variation (Gaffney & Keightley 2005), 2005) and that it may be functionally associated with chromsome arhictecture (Yaffe *et al.* 2010; Don *et al.* 2013). However this validity of this assertion is by no means definitive, as will be explored in chapter 5.

1.3.2.3 Small scale regional variation

Between 2 and maybe 2,000 bases, different features associate with mutation rates. Here we start to see the effects of CpG islands, these are ~1kb regions with increased density CpG dinucleotides (C immediately 5' of a G) and overall increased GC content relative to the genomic average (Gardiner-

Garden & Frommer 1987). C followed by G is often methylated in mammals, and methylated cytosine undergoes a high rate of deamination to generate T resulting in a C>T mutation (Bulmer 1986; Cooper & Krawczak 1990; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000). This is the most hypermutable point mutation in the human genome (Hwang & Green 2004; Nachman & Crowell 2000), however as CpG dinucleotides are usually hypomethylated in CpG islands, so the rate of CpG>TpG decreases whilst the rate of other point mutation remains unaffected (Tomso & Bell 2003; Cohen *et al.* 2011). There is however a potential role for selection maintaining low-level CpG methylation in CpG islands (Panchin *et al.* 2016).

DNase hypersensitivity, intimately linked to nucleosome movement which controls chromatin dynamics (Hughes & Rando 2014; Gross & Garrard 1988), is another feature that appears to have an effect on the mutation rate, reducing the rate of ~200bp regions, particularly strongly for CpG > TpG transitions (Ying *et al.* 2010). Transcription factors bind at DNase hypersensitive sites to initiate a process that itself has mutational biases; Where the non-transcribed strand, likely due to the actions of more efficient transcription coupled repair on the transcribed strand, has an excess of C>T transitions (Mugal *et al.* 2009; Polak & Arndt 2008; Beletskii & Bhagwat 1996; Green *et al.* 2003). However this small bias in the spectrum of mutation has very little effect on the overall mutation rate of transcribed sequences (Polak & Arndt 2008; Green *et al.* 2003; Hodgkinson & Eyre-Walker 2011) .

Multi-nucleotide mutations (MNMs) are complex mechanisms that produce two or more closely spaced mutations and the genomic distance over which these mechanisms act generally range from 2 to 100 bases (Schrider *et al.* 2013; Schrider, Hourmozdi & Hahn, 2011; Averof *et al.* 2000; Smith *et al.* 2003; Harris & Nielsen 2014; Terekhanova *et al.* 2013; Rosenfeld *et al.* 2010). It is estimated that 1-2% of all human single nucleotide polymorphism (SNPs) (Rosenfeld *et al.* 2010; Schrider,

Hourmozdi & Hahn, 2011; Harris & Nielsen 2014) and up to 3% of substitutions between primates (Schrider, Hourmozdi & Hahn, 2011; Terekhanova *et al.* 2013) are caused by MNMs, ~50% of which are at adjacent nucleotides (Smith *et al.* 2003; Harris & Nielsen 2014; Seplyarskiy *et al.* 2014; Chen *et al.* 2014). It is obvious that not accounting for these effects would inflate the apparent heterogeneity in mutation rates across the genome. This is difficult however as very little is known about the causes or how to suitably control for these effects. Although evolutionary processes could be responsible for MNMs, the evidence seems to favour a mutational origin for most MNMs (Schrider, Hourmozdi & Hahn, 2011; Averof *et al.* 2000; Smith *et al.* 2003; Harris & Nielsen 2014; Terekhanova *et al.* 2013). However the only MNM so far with robust evidence for a causative mechanism is GC>TT/AA. This mutation is a major signature of the error prone DNA polymerase zeta (Stone, Lujan & Kunkel, 2012; Harris & Nielsen, 2014; Seplyarskiy, Bazykin & Soldatov, 2014) accounting for 27% of all MNMs and thus responsible for a significant number of SNPs segregating in the human population (Harris & Nielsen 2014). Other emerging research from *Homo sapien-Pan troglodytes* alignments also attributes a significant proportion of MNMs separated by longer genomic distances to short template switching events (Löytynoja & Goldman 2016).

A further type of clustered mutation is result of overactive AID/APOBEC family of cytidine deaminses. These cause multiple clustered mutations spanning ~ 50bps with a preference for C>T/G (Pinto *et al.* 2016; Seplyarskiy *et al.* 2016). Although this mechanism was first discovered in somatic tissues (Nik-Zainal *et al.* 2012; Alexandrov *et al.* 2013) where it potentially affects much larger regions, it has recently been shown to have played an important role in evolutionary dynamics, at least among the hominids (Pinto *et al.* 2016), displaying a preference for the lagging strand during DNA replication and potentially accounting for 20% of C>T/G human SNPs (Seplyarskiy *et al.* 2016). To what degree this could confound results based upon the independence

of point mutations is currently unclear, but is emerging as something that needs to be considered when analysing substitution and SNP data.

1.3.2.4 Single nucleotide scale

The greatest variation in the mutation rate is found at the single nucleotide level (Hodgkinson & Eyre-Walker 2011). In part this variation is due to context - the identity of the nucleotides surrounding a site (Bulmer 1986; Cooper & Krawczak 1990; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000). The most well known example of a context effect is the previously mentioned CpG deamination where methylated C undergoes a high rate of deamination to generate T (Bulmer 1986; Cooper & Krawczak 1990; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000). It has been estimated that CpGs undergo rates of mutation 10-15 fold higher than other sites in the human genome (Hwang & Green 2004; Nachman & Crowell 2000) and generate ~20% of all mutations (Fryxell & Moon 2005). There are also other context effects, but these lead to variation in the mutation rate of only 2 to 3-fold (Hwang & Green, 2004; see Fig. 1a in (Hodgkinson & Eyre-Walker, 2011)), and the mechanism of this variation is poorly understood.

There also appears to be variation in the mutation rate at the single nucleotide level that does not depend upon simple context, variation that has been termed "cryptic variation in the mutation rate". This variation is cryptic in the sense that it can be detected within currently available mutation and substitution data and we can quantify the magnitude of it, but we currently have no explanation for its cause (Hodgkinson & Eyre-Walker 2011). The evidence for this variation comes from comparisons of various primates, including *Homo sapiens*, that there are an excess of sites in which

different species share a SNP, even when the influence of context on the mutation rate is taken into account. These excess SNPs could be due to sequencing error, assembly error of paralogous duplications or ancestral polymorphism. However, the evidence is in favour of a mutational mechanism that is not associated with simple context (Hodgkinson *et al.* 2009; Johnson & Hellmann 2011). It has been estimated that cryptic variation may generate more variance in the mutation rate than simple contexts, such as the CpG effect (Hodgkinson *et al.* 2009; Johnson & Hellmann 2011).

1.3.3 Direct Methods.

More direct approaches, using DNMs from trio data (Hodgkinson *et al.* 2009), enable the elucidation of germ-line mutational patterns without the noise of evolutionary processes, such as selection. Although rapidly increasing, the number of currently available whole genome DNMs for analysis are small (~43,000) when compared to the number of substitutions between *Homo sapiens* and *Pan troglodytes* whole genomes, and the size of the *Homo sapiens* genome. However, one thing that is starting to emerge, is that DNM rate variation, at least at large scales, may be less than previously suggested by patterns of substitutions from comparative studies (Michaelson *et al.* 2012; Kong *et al.* 2012; Wong *et al.* 2016; Francioli *et al.* 2014). In one of the first trio studies to compare whole genomes of *Homo sapiens*, Michaelson *et al.* (Michaelson *et al.* 2012) only find significant effects on site mutability for GC content at the 1Mb scale, recombination rate at the 100kb scale, nucleosome occupancy and DNase hyper-sensitivity at the 100bp scale and simple nucleotide context at the 1bp scale (Michaelson *et al.* 2012), and out of these simple nucleotide context has by far the strongest association. All other associations from this study appear insignificant, however their influence can not be fully disregarded; this study could be considered to be under-powered,

with only 581 DNMs from which to draw conclusions. In chapter 5 we seek to investigate these relationships further with the greater power enabled by the ~75 fold increase in the number available DNMs.

Probably the most noticeable difference between patterns of DNMs and substitutions concerns replication time. This shows a strong relationship with increased substitutions and SNPs in later replicating regions (Michaelson *et al.* 2012), but with respect to DNM density appears to be either insignificant (Stamatoyannopoulos *et al.* 2009; Chen *et al.* 2010) or has a small, converse relationship with paternal age; The number DNMs not only increase with paternal age (O'Roak *et al.* 2012; O'Roak *et al.* 2011; Iossifov *et al.* 2012; Iossifov *et al.* 2014; Kong *et al.* 2012; Neale *et al.* 2012; Michaelson *et al.* 2012), but are also overrepresented in early replicating regions (Francioli *et al.* 2015).

The most Robust evidence of mutation rate variation from direct studies is the quantification of the relative rates of mutations for four major classifications of DNMs; According to Kong *et al.* (Kong *et al.* 2012), the hyper-mutable CpG transitions (CTs) driven by spontaneous deamination (Bird 1980) occur 30 fold more often than the least mutable group, non-CpG transversions (NTv), with CpG transversions (CTv) and non-CpG transitions (NTs) occurring 2-3 fold more frequently than NTv. Interestingly, the data from Michaelson *et al.* (Michaelson *et al.* 2012) only estimates the rate for CTs to be ~20 fold greater than NTv, highlighting possible procedural differences associated with NGS technologies and pipelines or maybe reflecting ascertainment biases (Eyre-Walker & Eyre-Walker 2014). Similar values were also inferred over a decade earlier from comparisons of pseudogenes (Nachman & Crowell 2000), suggesting that inferences from established comparative methods represent the overall spectrum of DNMs relatively well, despite the apparent lack of agreement regarding associations with genomic features like replication time.

Direct methods have confirmed that idea of a mutational origin for MNMs using human trio data, estimating between 2.1% to 6.9% of all DNMs are MNMs, the lower bounds of which better agree with the estimates from comparative methods (Schrider, Hourmozdi & Hahn, 2011). Direct methods, in the form of mutation accumulation experiments, have also extended the search for MNMs to other organisms. These have produced similar estimates of between 2.8% and 3.4% for species as diverse as *Chlamydomonas reinhardtii* (Ness *et al.* 2015), *Drosophila melanogaster*, *Caenorhabditis elegans, Arabidopsis thaliana and Saccharomyces cerevisiae* (Schrider, Hourmozdi & Hahn, 2011; Schrider *et al.*, 2013). It is starting to become clear that MNMs are ubiquitous across taxa, and through violating the assumption of mutational independence, have the potential to not only confound studies of mutation rate heterogeneity, but also to enable unique evolutionary mechanisms. For example, the GC>TT MNM described previously makes it possible for one mutational event to convert the codon GCU (alanine) to UUU (phenylalanine), a change of amino acids that is not possible with just one point mutation.

It is clear that the much of information we have about mutation rate variation has been inferred from comparative studies. It is possible that the importance of this variation to DNM rates has been over-interpreted. This may be particularly pertinent at larger scales. To address this further, analysis of mutation rates using direct methods is required, and this is where chapters 2 and 5 of this thesis aims to contribute. Chapter 2 provides the first quantitative analysis of cryptic variation using direct methods, whilst chapter 5 examines, the relationship between DNMs, substitutions, SNPs and genomic features at different scales.

1.4 Somatic.

1.4.1 The overall rate of somatic mutation.

Prior to NGS technology, Information about the somatic rate of mutation was only known from calculations from cell line studies at single loci, like the HPRT (Bhattacharyya *et al.*, 1995; Lichtenauer-Kaligis *et al.*, 1996; Glaab & Tindall, 1997; Herrero-Jimenez *et al.*, 2000) and PIG-A genes (Araten *et al.*, 2005), or from monogenic somatic disease prevalence such as unilateral retinoblastoma caused by mutations in RB1 (Hethcote & Knudson, 1978; Fitzgerald, Stewart & Suckling, 1983) and colon cancer caused by mutations in APC (Mizunol, Watanabe & Iwama, 1993; Iwama, 2001; Luebeck & Moolgavkar, 2003; Hornsby, Page & Tomlinson, 2008). These studies calculated the somatic mutation rate to be 10-1000 fold higher than the germ line (Lynch, 2010) with the higher frequencies largely dependent on mismatch repair (MMR) deficiency (Mizunol, Watanabe & Iwama, 1993; Glaab & Tindall, 1997; Iwama, 2001; Luebeck & Moolgavkar, 2003; Hornsby, Page & Tomlinson, 2008). Some studies in HPRT (Lichtenauer-Kaligis *et al.*, 1996) did uncover variation in the mutation rate dependent upon genomic location, however no reason could be given for this.

With the application of NGS to whole cancer genomes, huge numbers of somatic single nucleotide variants (SNVs) quickly became available for analysis. It soon became clear that the mutation rate varied massively both within and between cancer types (Lawrence *et al.*, 2013) and that the overall mutation rate of a cancer explained 50% of the lifetime risk of developing cancer (Hao, Wang & Di, 2016). Further to this, different cancers displayed different mutational "signatures", thought to be indicative of the underlying genomic instability (Nik-Zainal *et al.*, 2012, 2016; Alexandrov *et al.*,

2013; Morganella *et al.*, 2016). However, it could be argued that adequate controls of non-cancerous whole somatic genomes are missing. Somatic SNVs are generally identified as such if they are present at sufficient frequency in DNA from tumour samples, but absent in the peripheral blood lymphocytes (a proxy for the germ-line) of the same individual. But despite thousands of whole cancer genomes being made freely available through portals such as ICGC (Zhang *et al.*, 2011), very few non-cancerous somatic cells have been sequenced. It is known that there is much diversity in large scale genomic aberrations of healthy somatic cells; copy number variations (CNV) and chromosomal ploidy changes have been reported (Yurov, 2005; O'Huallachain *et al.*, 2012). Healthy tissues exposed to ultra-violet light carry 2-6 DNMs per Mb, similar to cancerous tissues, 20% of which harbour driver mutations (Martinocorena & Campbell, 2015). A further study in healthy somatic tissues from *Mus musculus* suggested a bias for C>T, C>A and G>T in different tissues, highlighting the potential diversity of the somatic mutational spectrum dependent upon cell type, age and exposure to mutagens (Behjati *et al.*, 2014). Thus even though we now have millions of somatic SNVs obtained in a direct manner from whole genome sequenced cancers, it is not clear to what extent there is variation in non-diseased tissues and to what extent the patterns found in cancer genomes represent neutral variation or clonal selection during tumour development.

1.4.2 Somatic mutation rate variation across the genome

1.4.2.1 Whole Chromosomes.

Contrary to the germ-line, chromosomes in somatic cells spend the entirety of the time in just one gender, with somatic tissues from both sexes experiencing roughly similar numbers of cell divisions. Thus the driving forces behind mutation rate variation in the autosomes and the X/Y

chromosomes should no longer be applicable. The little research that has been done on this subject suggests that the X chromosome has a density of mutations roughly similar to the autosomes (Hodgkinson, Chen & Eyre-Walker, 2012); the Y chromosome data was not available in this study.

1.4.2.2 Large scale.

Efforts to understand mutation rate heterogeneity in somatic SNVs, like germ-line studies, has been primarily carried out at the Mb scale (Hodgkinson, Chen & Eyre-Walker, 2012; Schuster-Bockler & Lehner, 2012; Lawrence et al., 2013; Liu, De & Michor, 2013; Polak et al., 2015; Supek & Lehner, 2015) or 100kb scale (Koren et al., 2012). These have revealed many similarities to germ-line determinants from comparative studies. One of the first study showed that at the megabase scale, replication time, distance to telomere, GC content, gene density and nucleosome occupancy explained about 40% of the variance in mutation density, but each factor only explains very little of this variance (Hodgkinson, Chen & Eyre-Walker, 2012). Subsequent studies using considerably more genomes and investigating many more genomic variables claim to have extended the explainable variance to between 55% (Schuster-Bockler & Lehner, 2012) and 86% (Makova & Hardison, 2015; Polak et al., 2015), attributing most of the variance to histone modifications - particularly H3K9me3 - responsible for chromatin organisation. Chromatin architecture in general has been cited as the major determinant of large scales somatic mutation rate variation, where regions of closed chromatin, associated with histone acetylation in the form of H3K27, harbour higher mutation rates with open chromatin appearing less mutable, possibly due to easier access to DNA repair pathways (Schuster-Bockler & Lehner, 2012; Makova & Hardison, 2015). The aforementioned mutational increase in late replicating regions is also associated with an increase in transversions over transitions and appears unequivocally stronger in somatic tissue (Woo & Li,

2012) than the germ-line, in which recombination rate appears to be the more dominant association (Duret & Arndt, 2008). Similar to the germ-line, the effect of transcription on mutation rate, although highly evident in somatic cells, appears to be weak (Koren et al., 2012; Hao, Wang & Di, 2016). Another form of mutation pertinent to cancer etiology is that of fragile sites (FS) (Debatisse et al., 2012; Ozeri-Galai, Bester & Kerem, 2012). These are large regions of genomic instability that are associated with late replication and overlap large tumour suppressor genes, resulting in double strand breaks that that potentially knock out the tumour suppressors (Wilson et al., 2015), however very little is known about the biological mechanism of induction or how this form of large scale genomic instability is associated with SNVs. Chapter 4 attempts to tackle this question by investigating the correlation of fragile sites with SNVs and replication time.

With the obvious goal of cancer genomics being to discover the causative drivers of tumorigenesis, it is apparent that not properly accounting for mutational heterogeneity will lead to spurious results. Thus methods have been developed to control for the background rate of mutation based upon these associations. One such method is MutSigCV which primarily accounts for large scale variation in replication time and has proven successful in eliminating mutations in olfactory genes which would appear to have no biological relevance to cancer (Lawrence et al., 2013). However the level of variation seen at this scale only varies by about 5-fold (Hodgkinson, Chen & Eyre-Walker, 2012; Lawrence et al., 2013; Makova & Hardison, 2015). It is not currently clear how this compares to the magnitude of variation at the 1Mb scale in the germ-line, but data from taking comparative methods currently suggest a more modest 2-3 fold variation and so it is quite possible that there is a large disconnect between the two. Indeed the determinants of somatic variation only moderately agrees with data from comparative germ-line methods (Hodgkinson & Eyre-Walker, 2011; Makova & Hardison, 2015). Further, this 5-fold variation is also a fraction of the level found in both the soma and the germ line at the single nucleotide scale and thus the relevance of these large scale

associations to the major driving forces of mutation rate heterogeneity is questionable - being able to explain 86% of a quantity that only varies 2-3 fold may not explain much.

1.4.2.3 Small scale regional

There is much less data on small scale regional variation in the somatic mutation rate, and what there is seems to mimic the germ-line quite well. The over-expression of APOBEC enzymes was first discovered in somatic tissues (Nik-Zainal et al., 2012, 2016; Alexandrov et al., 2013) and is the one mechanism where more is known about the soma than germ-line. It can cause clusters of TpC>TpT/TpG mutations, independently of replication time (Morganella et al., 2016), potentially spanning up to 200kb (Roberts et al., 2012). It has also been suggested that it can be triggered by human papilloma virus (HPV), although the mechanism of action has yet to be determined (Hollstein et al., 2016). In addition it has also been implicated in causing complex mutations by attacking single stranded DNA in palindromic sequences, causing the recurrent MNM T**G**AA**C**AA > T**A**AA**T**AA in the *PLEKHS1* promoter region (Nik-Zainal et al., 2016). Interestingly though, unlike the many studies in the germ-line, there is yet to be any study conducted into the proportion of somatic DNMs that are MNMs. Nucleosome occupancy, like the germ-line also reduces mutation rates, although it has been shown to be mutation signature specific, as are the somatic associations with replication time (Morganella et al., 2016).

1.4.2.4 Single nucleotide level

The high heterogeneity between cancers in exposure to both exogenous and endogenous mutagens, hypermutators and various form of DNA editing and repair result in huge differences in rates of

mutation (Nik-Zainal et al., 2016). For example smoking associated lung cancers will be smothered with C>A mutations, whereas melanomas will be enriched for the C>T or CC>TT mutations resulting from UVb radiation (Hollstein et al., 2016). Oxidative DNA damage which also result in C>A mutations (Mugal, von Grunberg & Peifer, 2009), requires subsequent repair by the nucleotide excision repair (NER) pathway (Melis, van Steeg & Luijten, 2013), which has its own mutational biases (Roberts & Gordenin, 2014). MMR deficient colon cancers will be enriched for both C>T and C>A mutations (Alexandrov et al., 2013) and so it is obvious not only how heterogeneous different cancers are but also how hard it is to distinguish the causes of this heterogeneity. Chapter 3 attempts the first single nucleotide scale quantification of cryptic variation in the somatic mutation rate, by controlling for nucleotide context, thus aiming to ascertain the non-random patterns of mutation. However, what transpired was a clear indication of the current limitations of NGS technology and thus a clear quantification is not currently possible. The best estimates of single nucleotide somatic mutation rates so far appears to suggest that just like in the germ-line, CpG deamination occurs in a relatively clock like manner (Alexandrov et al., 2015). However, where in the germ line CpG deamination would be expected to be clock like respective to absolute time it appears to fit better with the number of cell divisions in the soma (Alexandrov et al., 2015).

1.5 The question of Scale

As seen above, scale is important to consider as it can produce different results, 1Mb has been suggested as the correct scale at which to study mutation rate variation (Gaffney & Keightley, 2005). At this scale the mutation rate varies maximum 3 fold in divergence data and 5 fold in somatic data across the genome (Hodgkinson & Eyre-Walker, 2011; Hodgkinson, Chen & Eyre-Walker, 2012; Lawrence et al., 2013; Makova & Hardison, 2015). The only study of DNMs to look

at the variation at this scale does not provide much evidence for such strong variation (Michaelson et al., 2012). More recently it has been suggested that actually the single nucleotide level is most appropriate scale at which to consider mutation rates (Ness et al., 2015), where the variation is much greater, up to 30 fold due to simple nucleotide context (Kong et al., 2012; Michaelson et al., 2012) and perhaps greater if considering the evidence for cryptic variation (Hodgkinson, Ladoukakis & Eyre-Walker, 2009; Johnson & Hellmann, 2011) . It is possible that the patterns of variation seen at the Mb scale are just manifestations of processes that occur at much smaller scales. Chapter 5 investigates the relationship of divergence with DNMs and recombination rate and other genomic features at 1Mb, 100kb scales to try to further understand four particular points; (i) how does the variation in DNM rate differ between these two scales, (ii) how well does variation in divergence along the genome represent the variation in DNMs, (iii) How do these patterns change over evolutionary time, and (iv) how well do the often cited determinants of mutation rate represent the variation in DNM density along the genome.

1.6 The contribution of this thesis

1.6.1 Overview

In the germ-line, mutation rate variation has often been inferred from divergence data which can include other confounders, such as bias gene conversion and natural selection. Only recently has the number of DNMs from pedigrees become substantial enough to be able to directly quantify mutation rate variation. In the soma, a proliferation of whole genome sequenced cancers have also provided millions of somatic SNVs to be analysed, showing both similarities and differences to patterns in the germ-line inferred from divergence. This thesis aims to further quantify mutation rate

variation in the germ-line directly from DNMs, exploring the magnitude of this variation at multiple scales and drawing comparisons to features that are associated with variation in rates of divergence. Additionally, it lays the ground for further exploration of these themes in somatic tissue, for which there is an increasing abundance of data, but sometimes lacking accuracy and suitable controls.

1.6.2 Chapter 2, cryptic variation in the germ-line.

In chapter 2, I start by providing further evidence for cryptic variation in the mutation rate, something that has only previously been inferred from indirect methods. Using the distribution of recurrent DNMs in autosomal dominant disease genes, I show that at the single nucleotide level, cryptic variation exists and may be huge, potentially dwarfing that of simple context mutations, such as CpG deamination. This is the first direct evidence of cryptic variation in the mutation rate in the germ-line

1.6.3 Chapter 3, cryptic variation in cancers and sequencing error.

In chapter 3, I extend the methodology of chapter 2 to exploit the vast number of somatic SNVs produced from the application of NGS methods to whole cancer genomes. This is the first attempt to quantify cryptic variation at the single nucleotide level in somatic tissues. The large number of mutations, ~3 million, allowed an extension of the method to consider each triplet - the base at which the SNV occurs and the neighboring bases - individually. The level of heterogeneity present was huge with an excess of recurrently hit sites that, when modelled with an additional parameter, indicated a high level of sites specific error in NGS. With this level of error, quantifying the cryptic

variation was not possible, but this has laid the ground for future similar attempts when the error rate of sequencing technology improves and has the potential to produce a highly accurate picture of mutation rate heterogeneity at the single nucleotide level.

1.6.4 Chapter 4, fragile sites and somatic mutation rate variation.

In chapter 4, using the same SNV data as chapter 3, I investigate the relationship of fragile sites with SNVs and replication time. I show that there is currently no evidence for an association of fragile sites and SNVs, suggesting an independent biology for both forms of mutation. There is however huge potential for this approach to be improved in the near future. The rapid output of cancer data that we are currently witnessing means increased sample homogeneity and increased power to detect variation.

1.6.5 Chapter 5, Divergence, de novo mutations and the scale of variation.

In this final chapter, I return to the germ-line. Most of our understanding about variation in the mutation rate has been inferred using indirect methods involving, alignments and SNPs. Using a combination of alignments and the ~43,000 DNMs now available from pedigrees, I show that the degree to which variation in substitutions from alignments represents human DNMs is quite weak for most substitution types. I provide evidence that this lack of correlation between divergence and DNM density along the genome is likely due to the actions of GC-biased gene conversion, and investigate the differing relationships of genomic features, like replication time and nucleosome occupancy, with divergence and DNM density. I also propose that recombination is both mutagenic

and has other evolutionary effects, shown by its independent relationships with both DNM density and divergence. Furthermore these patterns appear to be relatively well conserved throughout the primates. Finally I provide evidence suggesting that the magnitude of mutation rate variation at the large scale is actually very small and thus the pertinent scale at which to study mutation rate variation is likely to be at smaller scales.

1.6.6 Summary.

In Summary, this work utilises the recent proliferation of data flowing from NGS technologies to confirm, for the first time using direct methods, previous findings relating to the determinants, magnitude and scale of variation in the human mutation rate. These findings concern both germ-line and somatic tissues, contributing not only to evolutionary biology, but also having a clinical relevance for both inherited diseases and cancer research. Thus there is potential widespread applicability of not only the findings, but also future implementations of the methods developed herein.

2. Extensive Variation in the Mutation Rate Between and Within Human Genes Associated with Mendelian Disease.

2.1 Abstract

We have investigated whether the mutation rate varies between genes and sites using *de novo* mutations (DNMs) from three genes associated with Mendelian diseases. We show that the relative frequency of mutations at CpG dinucleotides relative to non-CpG sites varies between genes and relative to the genomic average. In particular we show that the rate of transition mutation at CpG sites relative to the rate of non-CpG transversion is substantially higher in our disease genes than amongst DNMs in general; the rate of CpG transition can be several hundred-fold greater than the rate of non-CpG transversion. We also show that the mutation rate varies significantly between sites of a particular mutational type, such as non-CpG transversion, within a gene. We estimate that for all categories of sites, except CpG transitions, there is at least a 30-fold difference in the mutation rate between the 10% of sites with the highest and lowest mutation rates. However, our best estimate is that the mutation rate varies by several hundred-fold variation. We suggest that the presence of hypermutable sites may be one reason certain genes are associated with disease.

2.2 Introduction.

There is evidence that the mutation rate varies substantially across the human genome in the germ-line from studies of *de novo* mutations (DNMs) (Francioli *et al.* 2015; Michaelson *et al.* 2012) and from comparative genomics (reviewed in (Hodgkinson & Eyre-Walker 2011)). Although this occurs at a number of different scales the most dramatic variation is seen at the single nucleotide level (Hodgkinson & Eyre-Walker 2011). In part this variation is due to context - the identity of the nucleotides surrounding a site (Bulmer 1986; Cooper & Krawczak 1990; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000). The most well-known example of a context effect is that of CpGs; C followed by G is often methylated in mammals, and methylated cytosine undergoes a high rate of deamination to generate T (Bulmer 1986; Cooper & Krawczak 1990; Coulondre *et al.* 1978; Gojobori *et al.* 1982; Hwang & Green 2004; Nachman & Crowell 2000). It has been estimated that CpGs undergo rates of mutation 10-15 fold higher than other sites in the human genome (Hwang & Green 2004; Nachman & Crowell 2000) and generate ~20% of all mutations (Fryxell & Moon 2005). There are also other context effects, but these lead to variation in the mutation rate of only 2 to 3-fold (Hwang & Green 2004).

In addition to variation associated with context, there also appears to be variation at the single nucleotide level that does not depend upon the identity of the adjacent nucleotides, at least not in a simple manner, variation that has been termed cryptic (Hodgkinson *et al.* 2009). The evidence for this variation initially came from the observation that there is at least a 50% excess of sites in which humans and chimpanzees share a single nucleotide polymorphism (SNP), even when the influence of context on the mutation rate is taken into account (Hodgkinson *et al.* 2009). Such an excess could be due to sequencing error, assembly error of paralogous duplications or ancestral polymorphism. However, several lines of evidence suggest that these explanations do not explain the excess of

coincident SNPs. First, the distribution of allele frequencies amongst coincident SNPs is identical to non-coincident SNPs (Johnson & Hellmann 2011); if coincident SNPs were due to assembly errors or ancestral polymorphisms we would expect them to be more frequent in the population than other SNPs. Second, sequencing coverage is no greater at coincident SNPs than other sites (Johnson & Hellmann 2011). And third, there is also an excess of coincident SNPs between human and macaque (Hodgkinson *et al.* 2009), two species which are very unlikely to share ancestral polymorphisms. There is also an excess of sites with substitutions in two independent pairs of primate species (Johnson & Hellmann 2011). These lines of evidence therefore suggest that the excess of coincident SNPs most likely arises from variation in the mutation rate that is not associated with context, at least not sequence contexts that are close to the site in question. It has been estimated that cryptic variation may generate more variance in the mutation rate than simple contexts, such as the CpG effect (Hodgkinson *et al.* 2009).

Although, variation in the mutation rate is most conspicuous at a single nucleotide scale it has also been known for some time that the mutation varies at larger scales in the human genome (Matassi *et al.* 1999; Michaelson *et al.* 2012; Spencer *et al.* 2006). The scale of this variation remains poorly characterised but a recent analysis of where DNMs occur suggest that the variation is probably at a scale of 10,000s of base pairs (Michaelson *et al.* 2012). The variation in the rate of CpG and non-CpG mutations appears to be at least partly independent, because the variation correlates to different genomic variables (Tyekucheva *et al.* 2008), but no systematic analysis of the relative rates of CpG and non-CpG mutation has been performed to our knowledge.

Here we investigate two aspects of variation in the mutation rate. First, does the relative frequency of transition and transversion mutations at CpG and non-CpG sites differ between genes, and second, is there variation in the mutation rate for transition and transversion mutations within CpG

and non-CpG sites (e.g. does the rate of transition mutation differ between CpG sites within a single gene). We address these questions using a dataset of *de novo* mutations (DNMs) that have been discovered during clinical screening in three genes associated with Mendelian diseases. In each case the DNMs were discovered in an unbiased manner – the causative gene was sequenced in a patient with the disease and their parents who did not have the disease.

2.3 Methods

2.3.1 Data

DNMs were discovered as part of routine clinical screening for individuals suffering from bilateral retinoblastoma, neurofibromatosis type I and Rett's syndrome; these diseases are caused by mutations in *RB1*, *NF1* and *MECP2* respectively. All data were collected after Ethics committee approval at each of the institutions involved.

The *MECP2* data were gathered from RettBASE, International Rett Syndrome Foundation *MECP2* Variation Database (http://mecp2.chw.edu.au), a curated database for *MECP2* variants from research and clinical laboratories (Christodoulou *et al.* 2003). Variants included in this study were limited to those for which parental testing had been carried out, with both parents tested for female patients, or maternal testing for male patients, since the gene is X-linked. Only variants from studies in which exons 2-4 had been sequenced were included, and our analysis was restricted to this part of the gene.

The *NF1* data were gathered from the *NF1* LOVD database

(https://grenada.lumc.nl/LOVD2/mendelian_genes/home.php?select_db=NF1). Both parents were tested for the pathogenic mutation and the father was tested for paternity. We only included studies in which all exons had been sequenced in transcript variant 2, this differs from transcript variant 1 in missing exon 31 (formerly known as exon 23a).

The *RB1* data came from three laboratories. Mutations were identified using a number of approaches including sequencing, single strand conformational polymorphism, heteroduplex analysis and high resolution melt analysis. Mutations were confirmed in each case by direct sequencing. There is alternative start codon in *RB1* (Sanchez-Sanchez *et al.* 2007) so exon 1 was ignored in the analysis. Both parents were tested for all *RB1* variants. Some of the *RB1* data has been previously published (Price *et al.* 2014).

The transcript numbering that we use is from NM_004992.3 for *MECP2*, NM_000267 for *NF1* and NM_000321.2 for *RB1*. We focus our analysis on nonsense mutations since nonsense mutations are more likely to have consistent phenotypic effects (see results section for further discussion).

2.3.2 Testing for mutation rate variation between genes

We performed two tests of mutational rate heterogeneity. First we tested whether the relative rates of CpG transitions, CpG transversions, non-CpG transitions and non-CpG transversions were significantly different between genes and between the genes and the background rate. To do this we performed a chisquare goodness-of-fit test, in which we calculated the expected number of CpG transitions and transversions, and non-CpG transitions and transversions, assuming that the ratios between the various mutational types were the same in the two genes, by finding the parameters of a simple model which minmised the chi-square statistic. We assumed that each gene has its own

"mutation rate", which reflects both the intrinsic mutation rate and the probability that the mutation comes to clinical attention; let this be μ_i. If we assume that the relative rates of the different mutation categories are the same in two genes then without loss of generality we can let the rate of non-CpG transversions rate be μ_i and the rates of CpG transition, CpG transversion and non-CpG transition be $\mu_i r_{cts}$, $\mu_i r_{ctv}$, and $\mu_i r_{nts}$, where r_{cts}, r_{ctv} and r_{nts} are shared between the two genes. To test whether the patterns of mutation are the same we find the values of μ_i, r_{cts}, r_{ctv} and r_{nts} that minimise the chi-square value, comparing the observed and expected values. Having found the parameters that minimise the chi-square value we performed a goodness of fit test using the ch-square value with 3 degrees of freedom (because we have eight observations and we have estimated 5 parameters).

2.3.3 Testing for mutation rate variation within genes

Second, we tested whether the rate of mutation varied within a mutational category (e.g. CpG transitions). If the rate of mutation is the same across all sites of a particular type then DNMs should be randomly distributed across those sites. To test whether DNMs tend to recur at sites more often than by chance we generated the expected number of sites hit recurrently by DNMs by randomly distributing the observed number of DNMs of the required type (e.g. CpG transitions) across the sites of that type that could generate a nonsense mutation. For each randomized dataset we tabulated the number of sites a site was hit zero, once, twice…etc by a DNM. By repeating this randomization 10,000 times we derived the expected distribution of DNMs (i.e. the number of times a site is expected to have been hit by one, two…etc DNMs). We compared the observed to the expected using a chi-square test. However, the test statistic is unlikely to be chi-square distributed because some of the expected values can be very small, We therefore empirically determined the distribution of the chi-square statistic by calculating the chi-square statistic for each simulated

dataset using the expected values estimated across all simulated datasets (as we did for the observed data). We then compared the observed chi-square statistic to this distribution. The p-value was the proportion of simulated datasets that had a chi-square value greater than observed chi-square value plus half the simulated datasets that had an identical chi-square value; this latter condition prevents the test being overly conservative when there are few DNMs. We performed simulations to check that this method did not generate excessive levels of type I error. For a given number of DNMs and sites we randomly allocated DNMs across sites and tabulated the number of sites that had been hit 0,1,2…etc times. We then performed the analysis as though this was real data, and repeated this 1000 times for a given combination of sites and DNMs. Simulations confirm that it does not increase the level of type I error, although it can decrease it when there are very few DNMs (i.e. the test can be overly conservative).

To combine probabilities from the heterogeneity tests we used the unweighted z-method (Whitlock 2005); in this method we find the value of a normal distribution deviate, with a mean of zero and variance of one, that would yield the corresponding p-value – the z-value. These z-values can be added to yield a z-value with an expected value of zero and a variance equal to the number of tests that have been combined. The overall p-value is then obtained by converting the combined z-value back into a p-value. We set p-values in which no simulated data had a greater chi-square value to 0.0001, and no simulated data had a smaller chi-square to 0.9999.

2.3.4 Parameter estimation

"In modelling the variation in the mutation rate, various models have been used; early discrete rate-class models provided good approximations (Fitch & Margoliash, 1967) but became computationally unfeasible when more than 3 rate-classes were included (Yang, 1994). However, a

continuous model based on the gamma distribution, which is therefore not limited to just 3 rate-classes, has been shown to provide a good fit to many datasets (Golding, 1983). The gamma distribution, modeled with a mean of 1, can represent a huge variation of rates, from practically invarient where the shape parameter β tends towards infinity to extreme rate heterogeneity when β approaches zero, and although other continuous distributions such as the log-normal have been explored (Golding, 1983), the gamma distribution is an established choice (Yang, 1994). We assume that whilst the mutation rate is gamma distributed, the number of mutations in any one sequece is Poisson distributed; i.e. if the average mutation rate per site is $\alpha\mu$ in a particular sequence, where alpha is gamma distributed and μ is the mean mutation rate, then the observed number of mutations is Poisson distributed with an expected number of mutations is $l\alpha\mu$ where l is the number of nucleotides in the sequence. By assuming that the number of mutations is Poisson distributed weare assuming that they are rare and that they occur independently. Although, there is some evidence that some mutations occur concurrently, because there is an excess of mutations that occur within 10bp on the same haplotype, these mutations still consititute a very small fraction of the total (~3%) (Schrider, Hourmozdi & Hahn, 2011).

We estimated the variation in the mutation rate within a mutational category (e.g. CpG transitions) as follows. Let us assume that the mutation rate at a site is $\alpha\mu$ where μ is the mean mutation rate and α is a deviation from the mean that is taken from some distribution $D(\alpha)$, which has a mean of 1; in our analysis we assume that $D(\alpha;\beta)$ is a gamma distribution with a shape parameter β. The number of mutations at a site can be modeled as a Poisson process for the reasons described above, and hence the number of mutations at a site is Poisson distributed. The probability of observing x mutations at a site is therefore

$$G(x;u,\beta)=\int_0^{\infty} D(\alpha;\beta)\frac{e^{-u\alpha}(u\alpha)^x}{x!}d\alpha=\frac{1}{x!\Gamma(\beta)}\left(\frac{1}{\beta}\right)^{-\beta}u^x(\beta+u)^{-x-\beta}\Gamma(x+\beta) \quad (1)$$

which is the negative binomial distribution, where $u=\mu k$ and k is a parameter proportional to the chance of observing a DNM; this is dependent upon the incidence and interest in the disease. The number of sites with x mutations is multinomially distributed and hence the likelihood of observing n_x sites with x mutations is

$$L(u,\beta)=n!\prod_x\frac{G(x;u,\beta)^{n_x}}{n_x!} \quad (2)$$

where n is the total number of sites. We found the maximum likelihood values of the distribution using the Nelder-Mead algorithm as implemented in the `NMaximize` function in Mathematica (version 7). The model above is described for a single mutational category in a single gene. However, it is straightforward to expand the analysis across multiple mutational categories and genes. In each analysis each mutational category in each gene is allowed its own u parameter reflecting the fact that the chance of observing a mutation varies between genes, and that the rate of mutation varies between mutational categories. Confidence intervals on parameters were derived from the likelihood surface – i.e. by finding the parameter values that decreased the log-likelihood by 2 units.

2.4 Results

2.4.1 Data

We have analysed DNMs in three genes that are associated with Mendelian disease. The genes are *RB1*, mutations in which cause retinoblastoma; we only consider mutations causing bilateral retinoblastoma since this disease is almost exclusively caused by a *de novo* germ-line mutation, whereas unilateral retinoblastoma is usually caused by somatic mutations. The second gene we consider is *NF1*, mutations in which cause neurofibromatosis type I. And the third gene is *MECP2*, mutations in which cause Rett's syndrome.

It is critical to our analysis that all mutations in a gene have similar penetrance, otherwise any apparent variation in the mutation rate might be due to variation in penetrance (i.e. a site with multiple recurrent DNMs might have a high mutation rate or the mutation might be highly penetrant). As a consequence we only consider nonsense mutations, and in *RB1* and *NF1* we only consider sites at which nonsense mutations are predicted to be caught by nonsense mediated decay (NMD). Furthermore, in *RB1* we ignore data from the first exon because mutations in the first exon may display variable levels of penetrance due to alternative translation initiation sites (Sanchez-Sanchez *et al.* 2007). All the nonsense mutations we consider in *RB1* and *NF1* should therefore have the same probability of being detected. The analysis of *MECP2* is more complex because the vast majority of sites that could generate a nonsense mutation are in the last exon and hence would not be caught by NMD; hence some nonsense mutations, particularly those towards the end of the gene could be less penetrant than those earlier in the gene. Furthermore, it is possible that nonsense mutations in the second and third exons (first and second coding exons) are lethal and therefore not routinely observed. Figure 1 shows the distribution of DNMs along the *MECP2* gene. It is

conspicuous that almost all pathogenic mutations occur between the start of the final exon and the end of the transcription repression domain. As a consequence we analysed two datasets for *MECP2* – all sites at which a nonsense mutations could occur, and all sites at which nonsense mutations could occur between the first and last sites that have multiple DNMs (sites 423 to 889 inclusive). Reducing the dataset in this manner does not alter the relative rates of mutation greatly, but it does reduce the evidence for heterogeneity within mutational categories (see below); this reduced dataset can therefore be considered a conservative dataset.

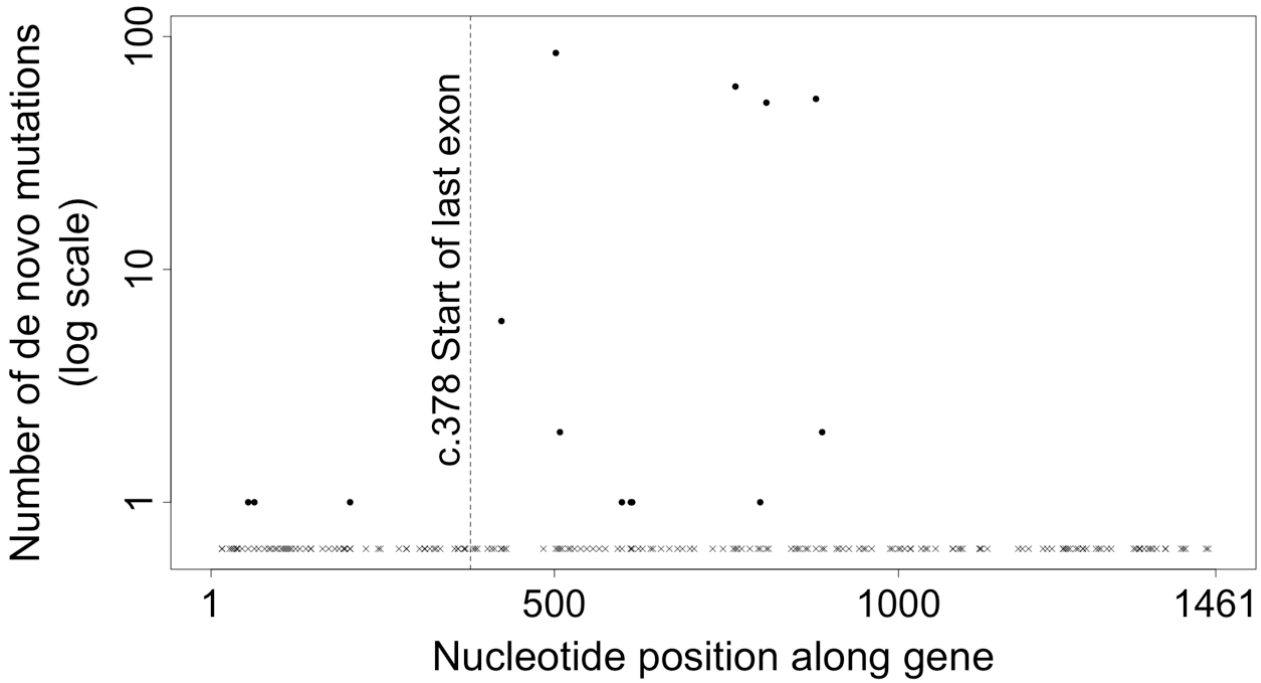


Figure 2.1. The distribution of nonsense DNMs and sites at which a mutation can cause a nonsense mutation in the MECP2 gene. The start of the last exon is marked. Crosses mark sites at which a mutation would generate a stop codon, filled circles mark sites at which nonsense DNMs have occurred and the number that have been observed.

Table 2.1 gives the number of DNMs in each of four mutation categories, transitions and transversions at CpG sites, and transitions and transversions at non-CpG sites, and Appendix 2.1 gives the number of sites hit by 0,1, 2…etc DNMs. We divided the data in this way because there are large differences in the rate of mutation of these mutational types (reviewed in (Hodgkinson & Eyre-Walker 2011). For each of our genes we have large numbers of nonsense DNMs. These are dominated by CpG transitions but we also have substantial numbers of non-CpG transitions and transversions.

	CpG ts	**CpG tv**	**non-CpG ts**	**non-CpG tv**
RB1	97	5	15	32
NF1	52	4	24	20
MECP2	253	0	6	12
MECP2 – restricted	252	0	2	10
Kong	855	73	2489	1516
Michaelson	70	10	282	185

Table 2.1. The numbers of nonsense DNMs in each gene and mutational category for three disease genes. The *MECP2* – restricted figures are for DNMs between positions 421 and 888. The Kong and Michaelson data are all the DNMs reported in Kong *et al.* (2012) and Michaelson *et al.* (2012) respectively.

2.4.2 Heterogeneity between genes

It is of interest to know whether the frequencies of different types of mutation vary substantially between genes. Unfortunately, because of the way in which our data have been sampled we cannot answer this question directly – the rate at which DNMs are detected in our genes depends upon the frequency of the disease, the severity of phenotype and the interest of clinicians. However, we can compare the relative frequency of different types of mutation between genes and compare those between genes and to the genomic average. We test for differences between genes, and between genes and the genomic average using a chi-square goodness of fit test, fitting a model in which we assume the relative rates of mutation in the four mutational categories are the same in the two genes (or genes and genome) (see the Materials and Methods section).

Two recent studies have obtained substantial numbers of DNMs from the complete genome sequencing of trios (Kong *et al.* 2012; Michaelson *et al.* 2012). Surprisingly the relative frequencies of the four mutation types differ significantly between these studies (Table 2.2) (Chi-square goodness of fit test, X^2=8, df=3, p=0.045). The difference seems to be largely a consequence of a higher relative rate of CpG transitions in the data of Kong *et al.* (Kong *et al.* 2012) compared to the data of Michaelson *et al.* (Michaelson *et al.* 2012) (28x the rate of non-CpG transversions versus 19x) (Table 2.2).

	CpG ts	CpG tv	non-CpG ts	non-CpG tv
RB1	90	5.7	3.5	1
NF1	120	14	5.5	1
MECP2	640	0	4.8	1
MECP2 – restricted	240	0	1.5	1
Kong *et al.*	28	2.4	1.6	1
Michaelson *et al.*	19	2.7	1.5	1

Table 2.2. The rates of mutation expressed relative to the rate of transversion at non-CpG sites.The rates are derived by dividing the numbers of DNMs by the number of sites (Supp. Table S1), and then dividing the rate by the rate for non-CpG transversions.

The reason for this discrepancy is not clear; it may be due to different ages amongst the two cohorts, or different biases in the sequencing methods, as other analyses seem to suggest (Eyre-Walker & Eyre-Walker 2014).

However, more striking than the difference between the two trio datasets are the differences in the relative rates of mutation between these datasets and the three genes for which we have DNMs (Table 2.2); each of the disease genes shows higher rates of mutation relative to the rate of non-CpG transversion than trio datasets. The most dramatic difference is the relative rate of CpG transition mutation in the *MECP2* gene where the mutation rate is estimated to be over 240x higher than the rate of non-CpG transversion (640x for the complete *MECP2* dataset and 236x for the restricted dataset). Pairwise comparisons show that the patterns of mutation are highly significantly different between MECP2 and the other two genes (Chi-square goodness of fit tests: $X^2 > 99$, df=6, $p<0.001$ in both cases), but not between *RB1* and *NF1*. The patterns are also highly significantly different between each of the three genes and both the datasets of Kong *et al.* and Michaelson *et al.* (Chi-square goodness of fit tests: $X^2 > 86$, df=3, $p<0.001$ in all cases). Unfortunately, it is not possible to say from these data whether the large relative rates are due to a low rate of mutation at non-CpG sites or a high rate at CpG sites.

2.4.3 Heterogeneity within genes

The analyses above show that the relative frequency of different types of mutation varies between genes. We can also test whether the rate of mutation within each of these mutational types varies between sites within a gene. Using a chi-square test of heterogeneity (deriving the null distribution by randomisation) we find highly significant evidence of heterogeneity over the entire dataset

whether we consider all sites in the *MECP2* gene or the restricted *MECP2* dataset ($p < 10^{-5}$) (Table 2.3). Surprisingly we find significant homogeneity, not heterogeneity, for CpG transitions sites in the *RB1* gene (i.e. mutations are more evenly distributed between sites than one would expect by chance alone). The data for this gene comes from three different labs. None of these datasets shows this excessive homogeneity individually and in fact the dataset from Barts Hospital shows marginally significant evidence of heterogeneity (p=0.081, null distribution derived by randomisation) (Appendix 2.2). The datasets are not significantly different to each other (X^2=23.6, df=20, p=0.26) (Appendix 2.3). It therefore remains unclear why the combination of the three datasets leads to significant homogeneity. It seems most likely that has arisen by chance.

	CpG ts	**CpG tv**	**non-CpG ts**	**non-CpG tv**
RB1	0.999	0.0629	0.0969	0.0412
NF1	0.0052	0.0344	0.1312	0.6017
MECP2	<0.0001	-	0.1281	<0.0001
MECP2 – restricted	<0.0001	-	0.0997	<0.0001
Overall – without *MECP2*	0.65	0.0089	0.044	0.15
Overall (row above)		0.0087		
Overall – *MECP2* all	0.032	0.0089	0.02	0.0013
Overall (row above)		2.0×10^{-6}		
Overall - *MECP2* restricted	0.032	0.0089	0.016	0.0013
Overall (row above)		1.6×10^{-6}		

Table 2.3. Testing for mutation rate heterogeneity within genes and mutational categories. The table gives the probability of observing the distribution of DNMs across sites under the null hypothesis that sites are equally mutable. Probabilities were combined using the unweighted Z-method (Whitlock 2005). The data from individual genes are combined in a number of different combinations – with and without the *MECP2* data, and with the restricted *MECP2* data. The probabilities are combined for each mutational type, but also across genes and mutational types.

The strongest evidence for heterogeneity comes from non-CpG transversions in the *MECP2* gene. In the restricted dataset there are 38 sites at which a non-CpG transversion will generate a nonsense mutation and there are 10 DNMs that have occurred at these sites. However, 6 of the DNMs have occurred at one site (site 423); all of these are C>G changes even though a C>A would also generate a stop codon.

Using the heterogeneity analysis we can identify 3 sites that have mutation rates that are significantly above background levels (Table 2.4). The mutation rate at these sites relative to all other sites, of the same mutational type, in the respective genes are given in Table 2.4. For two of these sites the mutation rates are only modestly above background levels; this reflects the power that we have to detect significantly hyper-mutable sites in CpG transition sites because we have more data than in other mutational categories. However, in *MECP2* we estimate that site 423, the site which has been hit by 6 non-CpG transversion DNMs has a mutation rate at least 150x (or 56x in the restricted dataset) higher than the background rate of non-CpG transversion in this gene. There is no obvious context effect associated with these sites (Table 2.4).

Gene	Mutation type	Rate	Context	hg19 coordinates
MECP2	CpG transition	2.0 (1.5)	CCCCTCCCGGCGAGAGCAGAA	chrX:153,296,777
MECP2	non-CpG transversion	150 (56)	TGATTGCGTACTTCGAAAAGG	chrX:153,296,856
NF1	CpG transition	4	TGTTGGAAGACGACCTTTTGA	chr17:29,588,751

Table 2.4. Significantly hypermutable sites. Numbers in parentheses in the rate column are the rate of the hypermutable site relative to the mean of all other sites; estimates for *MECP2* are using the restricted *MECP2* data. The nucleotide underlined in the context column is the hypermutable site.

2.4.4 Quantification

The estimates of the mutation rate at the sites with significantly elevated mutation rates are crude; one would expect that as more data accumulate so more sites will be found to be significantly hyper-mutable and hence the estimates of the rates will increase as sites are excluded from the background level. Therefore to better quantify the variation in the mutation rate we used maximum likelihood to fit a model in which mutation rates were distributed according to a gamma distribution. We fit several models in which the distribution of rates was shared (i) across all genes and mutational categories, (ii) across genes but varied between mutational categories, (iii) across mutational categories but varied between genes and (iv) finally a model in which every gene and mutational category combination had its own distribution. Using likelihood ratio tests we find the best supported model is one in which the gamma distribution is specific to a mutational category but is shared across genes (Appendix 2.4).

If we consider the shape parameter estimates for each mutational category it seems that CpG transitions have a much lower level of variation than the other mutational categories (higher the values of the shape parameter indicate lower levels of variation) (Table 2.5). In contrast the other three categories show substantial variation. To quantify this variation we calculated the ratio of the rates from the 90^{th} and 10^{th} percentiles. Whereas CpG transitions show just 1.4 fold variation between the upper and lower deciles the ratio for all the other categories is very substantial; for non-CpG transitions there is 36-fold variation but for CpG transversions and non-CpG transversions we infer more than a 1000-fold variation. However, the confidence intervals on these individual estimates are large and are also compatible with modest levels of heterogeneity; this is due a lack of data. We therefore combined data from CpG transversions, non-CpG transitions and non-CpG transversions.

Mutation type	Shape	Ratio of the rates of first and last deciles
cts	63 (13, infinity)	1.4 (2.1, 1.0)
ctv	0.39 (0.069, infinity)	550 (7.7×10^{13}, 1.0)
nts	0.81 (0.26, infinity)	36 (8000, 1.0)
ntv	0.24 (0.11, 0.64)	16,000 (6.1×10^{14}, 70)
ctv+ntv	0.26 (0.13, 0.64)	8000 (3.0×10^{7}, 70)
nts+ntv	0.39 (0.20, 0.91)	550 (93,000, 27)
ctv+nts+ntv	0.39 (0.21, 0.85)	550 (56,000, 32)

Table 2.5. Estimates of mutation rate heterogeneity within genes and mutational categories. Estimates of the shape parameter of the gamma distribution and the ratio of the upper and lower deciles of the distribution. 95% confidence intervals, as inferred from the likelihood surface, are given in brackets. cts – CpG transitions, ctv – CpG transversions, nts – non-CpG transitions, ntv – non-CpG transversions

Our estimate of the shape parameter is 0.39 (0.21, 0.85) and this corresponds to a ratio of deciles of 550 (i.e. the top 10% of sites mutate at least 550x faster than the bottom 10% of sites) with 95% CIs of 32x, 56,000x. In other words there appears to be very substantial variation in the mutation rate within each mutational category, with the exception of CpG transitions.

2.5 Discussion

We have provided evidence for two types of mutational heterogeneity. First, we have demonstrated that there is substantial variation in the relative rates of CpG and non-CpG mutations. The most conspicuous pattern is the very high rate of CpG transitions relative to non-CpG transversions. Whereas on average CpG dinucleotides undergo transition mutations between 18 and 30 fold the rate of non-CpG transversions, in our three Mendelian disease genes they undergo 90 to more than 200-fold higher rates of mutation. It is not possible to infer whether this is due to a low rate of non-CpG transversion or a high rate of CpG transition. Second, we have shown that there is significant heterogeneity in the mutation rate between sites within each mutational category. This is particularly evident for all categories other than CpG transitions; we estimate amongst the other categories that the mutation rate may vary by 100-fold or more.

Our conclusions are conditional on the assumption that all the mutations, which we have considered, both those that have occurred and those that could occur in a gene, are equally likely to be sampled. Variation in sampling might arise through three processes – variation in penetrance, alternative splicing and ascertainment bias.

In an attempt to ensure that all mutation were equally penetrant we restricted our analysis to

nonsense mutations, and in the case of *RB1* and *NF1*, to nonsense mutations that are predicted to be caught by NMD. In the case of *MECP2* most sites that could cause a nonsense mutation are found in the last exon and hence would not be caught by NMD. In attempt to ensure that all mutations in *MECP2* had similar penetrance we analysed the pattern of mutation both amongst all DNMs and amongst a subset of DNMs between the first and last recurrently hit sites. We have found similar patterns. If we remove *MECP2* from our analysis we still find evidence that the ratio of CpG to non-CpG mutations varies between the two disease genes and the background rate, and there is still significant heterogeneity in the mutation rate within a mutational category (Table 3). Never-the-less it is difficult to completely rule-out variation in the level of penetrance as an explanation for our results; if the variation in the density of DNMs is due to variation in penetrance then our results suggest that penetrance varies considerably between mutational categories and between sites within a gene.

The apparent variation in the mutation rate could also be due to ascertainment bias. Although we restricted our analysis to data that had come from studies in which the same part of the gene had been analysed it is possible that the causative mutation was not ascertained and these cases discarded. If some mutations are more likely to escape detection than others then it will appear as though there is mutation rate variation.

The variation could also potentially be due to alternative splicing (or alternative translation start sites) since nonsense mutations in constitutive exons might be more penetrant (or more lethal) than nonsense mutations in alternatively spliced exons. However, this seems an unlikely explanation for our results. *RB1* is known to have an alternative translation start site (Sanchez-Sanchez *et al.* 2007) and as a consequence exon 1 was removed from the analysis. There are two major splice forms of *MECP2* which differ in both their translation start site and the inclusion of exon 2 – variant

1 includes exon 2, in which translation starts, whereas variant 2 excludes exon 2 with translation starting in exon 1 (Kriaucionis & Bird 2004). We have analysed data mapped to variant 1, which differs from variant 2 in the first 26 bp. Hence the variation we observe is unlikely to be a consequence of alternative splicing since the vast majority of the data we have analysed comes from exons that are found in the major splice forms. There are multiple splice forms of *NF1*, although most of them yield products that are removed by NMD or result in highly truncated proteins (Barron & Lou 2011). We have used data that maps to transcript variant 2, which is one of two major splice forms. This differs from variant 1 in missing exon 31 (formerly exon 23a), a 21 amino acid exon, found in variant 1. So again it seems unlikely that alternative splicing can be responsible for our results in *NF1* because we have analysed data only from the exons present in both of the two major splice forms.

Another potential explanation for our results is positive selection in the germ-line. It has been found that some pathogenic mutations are advantageous within the male germ-line leading to an increased prevalence of diseases such as Apert's syndrome, which is caused by mutations in the gene *FGFR2* (Goriely & Wilkie 2012). None of the genes that we have studied are known to have mutations that are positively selected in the germ-line and it seems unlikely that the heterogeneity amongst nonsense mutations could be caused by this process, since all nonsense mutations are predicted to have the same or similar phenotypes.

The variation in the density of DNMs is therefore most likely due to variation in the mutation rate. It has previously been shown that mutation rates vary at a regional scale (reviewed in (Hodgkinson & Eyre-Walker 2011)). However, it has not been noted before that the relative rates of CpG and non-CpG mutation can vary substantially. The magnitude of the variation that we have observed might in part be due to the fact that genes with the highest mutation rates are those most likely to be

associated with disease, assuming that the high CpG to non-CpG mutation rate reflects a high CpG mutation rate and not a low non-CpG rate. There is some evidence for this effect; a recent model of the mutation rate at sites in the human genome, based on the analyses of DNMs and where they occur, predicts that disease genes have higher rates of mutation than non-disease genes (Michaelson *et al.* 2012).

It has also been noted that the mutation rate can vary within a mutational category because of context (Hodgkinson & Eyre-Walker 2011)however the effects within CpG or non-CpG categories have been inferred to be quite modest. For example, Hwang and Green (Hwang & Green 2004) estimated, using the divergence between primate species, that on average CCG, ACG, GCG and TCG mutate 24, 29, 18 and 23-fold faster than the genomic average. Context effects at non-CpG sites are also fairly modest, typically showing 2-3 fold variation when the immediately adjacent nucleotides are considered (Hwang & Green 2004), getting progressively weaker as sites further away from the focal site are considered (Zhao & Boerwinkle 2002). The level of variation within all mutational categories except CpG transitions seems to be considerably greater than this. The substantial variation in the mutation rate within each mutational category, except CpG transitions, is consistent with the cryptic variation in the mutation rate, which was first identified in nuclear DNA from the coincidence of SNPs in humans and chimpanzees (Hodgkinson *et al.* 2009; Johnson & Hellmann 2011). As we have found here, Hodgkinson *et al.* (Hodgkinson *et al.* 2009) estimated that there was more variation in the mutation rate within non-CpG sites, than within CpG-sites, and estimated that a gamma distribution with a shape parameter of 0.85 (0.83, 0.87) fitted the data at non-CpG sites. This is not significantly different to the estimate obtained here, 0.39 (0.20, 0.91).

The variation at CpG sites could potentially be a consequence of variation in methylation.

Methylated CpGs are expected to mutate faster than non-methylated CpGs due to the instability of methyl-cytosine (Coulondre *et al.* 1978; Bird 1980; Sved & Bird 1990). None of the sequences that we have analysed contain CpG islands, regions of the genome in which CpGs are not methylated. However, some of the variation may be due to residual variation in methylation.

In summary we have shown that there is significant variation in the mutation rate both between and within genes. Some of this variation might explain why these genes are associated with disease; they have high mutation rates, either overall or at specific sites that can cause disease, and this makes it more likely that pathogenic mutations will recur in the human population and cause disease.

3. Are sites with multiple single nucleotide variants in cancer genomes a consequence of drivers, hypermutable sites or sequencing errors?

3.1 Abstract

Across indepedent cancer genomes it has been observed that some sites have been recurrently hit by single nucleotide variants (SNVs). Such recurrently hit sites might be either i) drivers of cancer that are postively selected during oncogenesis, ii) due to mutation rate variation, or iii) due to sequencing and assembly errors. We have investigated the cause of recurrently hit sites in a dataset of >3 million SNVs from 507 complete cancer genome sequences. We find evidence that many sites have been hit significantly more often than one would expect by chance, even taking into account the effect of the adjacent nucleotides on the rate of mutation. We find that the density of these recurrently hit sites is higher in non-coding than coding DNA and hence conclude that most of them are unlikely to be drivers. We also find that most of them are found in parts of the genome that are not uniquely mappable and hence are likly to be due to mapping errors. In support of the error hypothesis, we find that recurently hit sites are not randomly distributed across sequences from different laboratories. We fit a model to the data in which the rate of mutation is constant across sites but the rate of error varies. This model suggests that ~4% of all SNVs are error in this dataset, but that the rate of error varies by thousands-of-fold between sites.

3.2 Introduction

There is currently huge interest in sequencing cancer genomes with a view to identifying the mutations in somatic tissues that lead to cancer, the so called “driver” mutations. Driver mutations are expected to cluster in particular genes or genomic regions, or to recur at particular sites in the genome, because only a limited number of mutations can cause cancer. For example, the driver mutations in the TERT promoter were identified because it had independently occurred in multiple cancers (Huang *et al.*, 2013). However, there are two other processes that can potentially lead to the repeated occurrence of an apparent somatic mutation at a site. First, it is known that the mutation rate varies across the genome at a number of different scales in both the germ-line and soma (Hodgkinson & Eyre-Walker, 2011; Hodgkinson, Chen & Eyre-Walker, 2012; Michaelson *et al.*, 2012; Francioli *et al.*, 2015). Sites with recurrent SNVs could simply be a consequence of sites with high rates of mutations. And second there is the potential for sequencing error. Although, the average rate of sequencing error is thought to be quite low it is evident that some types of sites, such as those in runs of nucleotides, are difficult to sequence accurately. Furthermore, since the genome contains many similar sequences it can often be difficult to map sequencing reads successfully (Treangen & Salzberg, 2013).

In the germ-line the density of point mutations varies at a number of different scales (Hodgkinson & Eyre-Walker, 2011). At the mega-base scale the mutation varies by about 2-fold, and ~50% of this variance can be explained by correlations with factors such as replication time, recombination rate and distance from telomeres (as reviewed in (Hodgkinson & Eyre-Walker 2011)). However the greatest variance, reportedly up to ~30-fold, has been found at the single nucleotide level (Hodgkinson, Chen & Eyre-Walker, 2012; Kong *et al.*, 2012; Michaelson *et al.*, 2012), whereby the nucleotide context, that is the identity of the bases immediately 5’ and 3’ of the mutated base, are

highly influential on the rate of mutation (Gojobori, Li & Graur, 1982; Bulmer, 1986; Cooper & Krawczak, 1990; Nachman & Crowell, 2000; Hwang & Green, 2004). The most well known example is that of CpG hyper-mutation (Bird, 1980), which is thought to account ~20% of all mutations in the human genome (Fryxell & Moon, 2005). However there is also variation at the single nucleotide level that cannot be ascribed to the effects of neighbouring nucleotides; this has been termed cryptic variation in the mutation rate and is thought to account for at least as much variation in the mutation rate as does simple context (Hodgkinson, Ladoukakis & Eyre-Walker, 2009; Eyre-Walker & Eyre-Walker, 2014, Johnson & Hellman, 2011, Smith *et al.*, 2016).

The somatic mutation rate is estimated to be at least an order of magnitude greater than that of the germ line (Lynch, 2010). It has been shown to vary between cancers (Lawrence *et al.* 2013) and different cancer types are known to vary in their relative contributions of different mutations to their overall mutational compositions (Alexandrov *et al.*, 2013). For a review see (Martinocorena & Campbell, 2015). The aforementioned correlates of variation that are found in the germ line are also apparent in the soma (Hodgkinson, Chen & Eyre-Walker, 2012; Schuster-Bockler & Lehner, 2012; Lawrence *et al.*, 2013; Liu, De & Michor, 2013), for example replication time correlates strongly with single nucleotide variant (SNV) density at the 1Mb base scale and can vary by up to 3-fold along the genome (Hodgkinson & Eyre-Walker, 2011; Woo & Li, 2012). However, as yet there has been no attempt to quantify the level of cryptic variation in the mutation rate at the single nucleotide level in the somatic genome. This is an important property to understand; for example a site which experiences a recurrence of SNVs across many cancer genomes would be of interest as a potential driver of cancer (Lawrence *et al.*, 2013), however, this site might simply be cryptically hypermutable (Hodgkinson, Ladoukakis & Eyre-Walker, 2009; Eyre-Walker & Eyre-Walker, 2014; Smith *et al.*, 2016). Here we examine the distribution of recurrent SNVs taken from 507 whole genome sequences made publicly available by Alexandrov *et al.* (2013) to investigate the level of

cryptic variation in the mutation rate for somatic tissues. We show that there is a large excess of sites that have been hit by recurrent SNVs. Since the density of these is greater in the non-coding, than the coding fraction of the genome, we conclude that most of them are unlikely to be drivers. We therefore investigate whether they are due to mutational heterogeneity or sequencing errors. In particular we investigate whether there might be cryptic variation in the mutation rate in cancer genomes. Unfortunately, the available evidence suggests that most sites with recurrent SNVs are likely to be due to sequencing error or errors in post-sequencing processing.

3.3 Methods

3.3.1 Genome and data filtering

The human genome (hg19/GRCh37) was masked to remove simple sequence repeats (SSR) as defined by Tandem Repeat Finder (Benson, 1999). The remaining regions were separated into three genomic fractions, consisting of 1,346,629,686 bp of non-coding transposable element DNA (TE), defined as LINEs, SINEs, LTRs and DNA transposons as identified by repeat masker (Smit *et al.* 1996), 1,322,985,768 bp of non-coding non-transposable element DNA (NTE), and 119,806,141 bp of exonic non-transposable element DNA (EX) defined by Ensemble (Flicek *et al.*, 2011). From the supplementary data of Alexandrov *et al.* (2013) we collated 3,382,737 single nucleotide variants (SNV), classified as “somatic-for-signature-analysis” (see (Alexandrov *et al.*, 2013) for SNV filtering methods). These can be downloaded from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/. These came from 507 whole genome sequenced cancers and represent 10 different cancer types and were reduced to 3,299,881 SNVs when excluding SNVs in simple sequence repeates (SSRs); 1,666,759 in TE and 1,535,069 in NTE and

98053 in EX.

3.3.2 Testing for mutation rate heterogeneity

We were interested in whether some sites have more SNVs than expected by chance. Since the mutation rate is affected by the identity of the neighbouring nucleotides we need to control for those effects. To do this we separated each SNV into one of 64 categories based upon the triplet to which it was the central base. This was reduced to 32 triplets when accounting for base complementarity with the pyrimidine (C/T) taken as the central base. If the total number of triplets of type *i* (e.g. C**T**C in the non-TE fraction) is l_i and the number SNVs at that triplet is m_i then the expected number of sites hit *x* times can be calculated using a Poisson distribution:

$$E_i(x) = l_i P(x, \mu_i) \qquad (1)$$

where $P(x, \mu_i) = \frac{e^{-\mu_i}\mu_i^x}{x!}$ is the Poisson distribution and $\mu_i = m_i/l_i$ is the mean number of SNVs per site, The expected number of sites with *x* SNVs across all triplets was calculated by summing the values of $P_i(x)$. Whether the observed distribution deviated from the expected was tested using a chisquare goodness of fit test; the chisquare value for the observed *n* sites hit *x* times was calculated for each hit category (0-7 hits). The individual probabilities of observing *n* sites being hit *x* times is derived from a Poisson distribution and is calculated for $n=0$ to $n=\max(n)$ for each hit category (0-7 hits). The total probability for observing the data for each hit category can then be calculated as 1-sum(individual probabilities for each hit category).

3.3.3 Model fitting

As well as testing whether there was significant hetereogeneity we were also interested in quantifying the level of variation. We fit two basic models. In the first we allowed the density of SNVs to follow a gamma distribution. Let the expected density of SNVs at a site be $\mu\alpha$ where μ is the mean density of SNVs for a particular triplet and α is the deviation from this mean. This deviation is assumed to be gamma distributed, $D(\alpha;\beta)$, parametrized with a shape parameter of beta and a mean of one. Under this model the expected number of sites with x SNVs is

$$E_i(x;\beta)=l_i\int_0^\infty D(\alpha;\beta)P(x,\alpha\mu_i)\,d\alpha \qquad (2)$$

In a second model we imagine that the production of SNVs depends upon two processes, one of which is constant across sites, and one which varies across sites, with the rate, α, drawn from a gamma distribution. Let the proportion of SNVs due to the first process be ε. Then the rate at which SNVs occur at a particular site where the rate of α *is* $\mu_i\varepsilon + \alpha\mu_i(1-\varepsilon)$. Hence the expected number of sites with x SNVs is

$$E_i(x;\beta,\varepsilon)=l_i\int_0^\infty D(\alpha;\beta)P(x,\mu_i(\varepsilon+\alpha(1-\varepsilon)))\,d\alpha \qquad (3)$$

We can now sum across triplets to find the expected number of sites with x SNVs.

$$E(x)=\sum_{all\,i} E_i(x;\beta,\varepsilon) \qquad (4)$$

Since sites with SNVs are rare, the number of sites with x SNVs is approximately Poisson

distributed. Hence if we observe $\widehat{E(x)}$ sites with x SNVs the likelihood can be written as

$$L=\prod_{x=1}^{\infty} P(\widehat{E(x)}, E(x)) \qquad (5)$$

These likelihoods can be multiplied across triplets to obtain the overall likelihood. We estimated the maximum likelihood values of the model parameters using the `NMaximize` function of Mathematica (version 7) which implements the Nelder-Mead algorithm (Nelder *et al.*, 1965). Chisquare goodness of fit test, as described in *section 2.3.3* was used to compare the fit expected number of sites with *x* SNVs obtained by the models to the observed data.

3.3.4 Privacy analysis

To investigate whether the SNVs at some sites tended to be produced by a particular research group we took all sites with 3 or more SNVs from the same cancer type and then performed Fishers exact test on a 2 x 30 matrix using the the R stats package, version 3.2.4 (R Core Team, 2016).

3.3.5 Mappability

Each nucleotide in genome was assigned a mappability score for uniqueness, as determined by the Mappability track (Derrien *et al.*, 2012) downloaded from the UCSC table browser at http://genome.ucsc.edu/ (Karolchik *et al.*, 2004). This feature assigns a value of 1 to unique *k*-mer sequences in the genome, 0.5 to those that occur twice, 0.33 to those that occur thrice etc. This is computed for every base in the human genome with the value being assigned to the first position of

the k-mer. We used *k*-mers of 100 and 20 bases.

3.4 Results

3.4.1 The distribution of recurrent SNVs

If there is no variation in the density of single nucleotide variants (SNVs) then we should find them to be distributed randomly across the genome. To investigate whether this was the case we calculated the expected number of sites with 1,2,3…etc SNVs, taking into account the fact that some triplets have higher mutation rates than others. We found that there are some sites that have 7 SNVs whereas we expect very few sites to have more than 3 SNVs (Table 3.1A, Figure 3.1) – the difference is highly significant using the Chi-square goodness of fit test ($p < 0.0001$) for both the whole genome (Total) and when separating the genome into non-coding transposable elements (TE), non-coding non-transposable elements and (NTE) and exons (EX) (Table 3.1A). We refer to sites with 3 or more SNVs as excess sites. In total we observed 1187 excess sites (Table 3.1A) with the density of excess sites in TE being 3.9 and 3.4 fold greater than in NTE and EX respectively. The probability of this level of SNV recurrence by chance alone is so low (Chi-squared goodness of fit test, $p < 0.0001$) that these excess sites must either be (i) drivers, (ii) the result of mutation rate heterogeneity across the genome or, (iii) the consequence of next generation sequencing (NGS) pipeline errors.

A) – All Sites

Site Type	0 hits	1 hit	2 hits	3 hits	4hits	5hits	6hits	7hits
Non-Exon TE obs (TE)	1344972042	1649680	7034	762	130	26	9	3
Non-Exon TE exp (TE)	1344964359	1663896	1430	1.14	9E-04	7E-07	5E-10	4E-13
Non-Exon Non-TE obs (NTE)	1321454397	1527967	3171	188	35	6	2	2
Non-Exon Non-TE exp (NTE)	1321451907	1532655	1206	0.86	6E-04	4E-07	3E-10	2E-13
Exon obs (EX)	119708384	97488	245	23	0	0	1	0
Exon exp (EX)	119708145	97939	57	0.03	2E-05	7E-09	3E-12	1E-15
Total obs	2786134823	3275135	10450	973	165	32	12	5
Total exp	2786124411	3294490	2692	2.04	2E-03	1E-06	8E-10	5E-13

B) – Mappable 100

Site Type	0 hits	1 hit	2 hits	3 hits	4hits	5hits	6hits	7hits
Non-Exon TE obs (TE)	1223239922	1517676	3927	266	25	11	5	1
Non-Exon TE exp (TE)	1223236637	1523873	1322	1.07	9E-04	7E-07	5E-10	4E-13
Non-Exon Non-TE obs (NTE)	1276165087	1499761	2698	97	16	2	0	1
Non-Exon Non-TE exp (NTE)	1276163336	1503124	1201	0.88	6E-04	5E-07	3E-10	2E-13
Exon obs (EX)	112360615	93084	185	16	0	0	0	0
Exon exp (EX)	112360453	93392	55	0.03	2E-05	7E-09	3E-12	1E-15
Total obs	2611765624	3110521	6810	379	41	13	5	2
Total exp	2611760426	3120389	2578	2	2E-03	1E-06	8E-10	6E-13

C) – Mappable 20

Site Type	0 hits	1 hit	2 hits	3 hits	4hits	5hits	6hits	7hits
Non-Exon TE obs (TE)	388613299	480820	741	9	0	0	0	0
Non-Exon TE exp (TE)	388612958	481494	417	0.34	3E-04	2E-07	2E-10	1E-13
Non-Exon Non-TE obs (NTE)	892370709	1061716	1621	31	4	1	0	1
Non-Exon Non-TE exp (NTE)	892369874	1063340	868	0.65	5E-04	3E-07	2E-10	2E-13
Exon obs (EX)	74735962	61034	103	6.00	0	0	0	0
Exon exp (EX)	74735883	61187	36	0.02	9E-06	4E-09	2E-12	7E-16
Total obs	1355719970	1603570	2465	46	4	1	0	1
Total exp	1355718714	1606021	1321	1	8E-04	6E-07	4E-10	3E-13

Table 3.1. Observed and expected values for the distribution of SNVs for sites hit from 0-7 times. A) shows data for the whole interrogable human genome, excluding simple sequence repeats. B) shows data for all bases in the genome that are uniquely mappable at 100 base pairs. C) the same as B but for 20 base pairs. $P < 0.001$ for observing >7 sites with 3 SNVs in A),B) and C).

It seems unlikely that the majority of the excess sites are due to drivers since the density of excess sites is higher in the TE and NTE parts of the genome than in EX (Table 3.1A). Furthermore, to date only one intergenic driver of cancer – an activating C>T mutation in the *TERT* promoter (Huang *et al.* 2013) at chr5:1,295,228 – has been confirmed, and although this is included in the excess sites with 7 SNVs, the remaining 1186 excess sites are unlikely to be under such selection. It therefore seems likely that the excess sites are either due to mutation rate variation or problems with sequencing.

3.4.2 Excess sites are enriched in non-unique sequences

The human genome contains many duplicated sequences particularly within transposable elements, and these pose challenges for accurate alignment of the short ~100bp reads produced from NGS (Zhuang *et al.*, 2014). If the excess sites were the result of NGS mapping errors then we might expect them to occur in regions of the genome that were hard to align. Using the mappability scores (Derrien *et al.*, 2012) we excluded all bases that were not uniquely mappable at 100bp; this should give an overall indication of how easy it is to map reads to the region. This only reduced the interrogable genome by 6%, but the number of excess sites was reduced by 64% (Table 3.1B), demonstrating that a large proportion of the excess sites were in duplicated sequences and therefore likely originate from mapping errors. However, even with this large reduction in excess sites we still observed many excess sites far greater than chance expectation (Chi-squared goodness of fit test, $p < 0.0001$) (Table 3.1B & Figure 3.1).

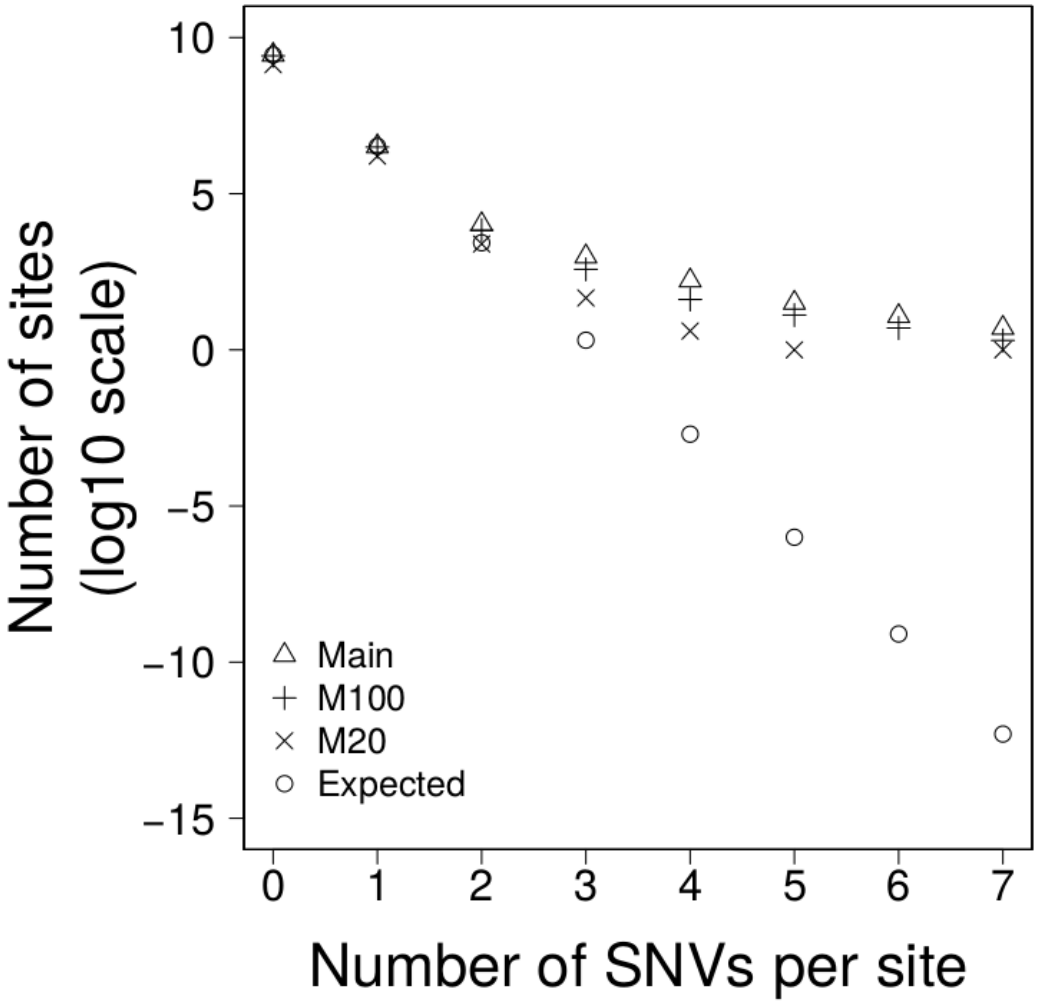


Figure 3.1. The number of sites with 0-7 SNVs per sites for: **Main** = all data, **M100** = sites that are uniquely mappable at 100 base-pairs, **M20** = sites that are uniquely mappable at base-pairs and, **Expected** is the expected number of SNVs per site drawn from a poisson distribution using all data.

The SNVs in this data were all called from >100bp reads. If the excess sites were errors of read mapping, they should not be affected by the uniqueness of shorter sequences (i.e. there is no reason why 100bp sequences that map uniquely to the genome should be mis-mapped if it contains a non-unique 20bp sequence), however if the SNVs were the product of a biological process that was more prevalent in non-unique or repetitive sequences, then we might expect to see a reduction of excess sites when we exclude all bases that do not map uniquely at 20bp. When we excluded all bases that were not unique at 20bp we found that the interrogable genome was reduced by 52% and the excess sites were reduced by 96% (Table 3.1C & Figure 3.1). It is worth noting that, due to their proliferative nature throughout the genome, this reduction disproportionately affects TEs where the interrogable genome is reduced by 71% and the excess sites by >99%. This would suggest that the excess sites existing in sequences that were unique at 100bp but not unique at 20bp may represent some biological process and not error. Furthermore, th*e TERT* promoter, whose recurrence is the result of positive selection, and is therefore the only excess site that that we can confidently say is not a product of error, remains in this most conservative of these analyses. Despite this large reduction in excess sites, significant heterogeneity still remains; the probability of observing the 52 excess sites in the part of the genome uniquely mappable at 20 bases is still extremely low (Chi-squared goodness of fit test, $p < 0.0001$).

One other potential problem with mapping reads to non-unique sequences occurs when a segmental duplication has been collapsed in the assembly of the reference genome; i.e. reads from two different locations are mapped to the same locus in the reference. Differences between the duplications will appear as SNVs. If this was the case we would expect to see an increase in excess site read coverage of ~2-fold or greater. To investigate whether this could be a problem in our data we compared the read coverage for excess sites and non-excess sites, which nevertheless had an SNV, in the one set of cancer genomes for which we had this information - the liver cancers

sequenced by the RIKEN group. However, we found that the median read coverage for the excess sites (n=15) was actually lower than for non-excess sites (n=224602) (28 and 33 reads respectively; Mann-Whitney U test, p = 0.043)

3.4.3 Privacy of mutations

To further investigate the origin of excess sites we exploited the fact that some types of cancer were sequenced by different laboratories using different technologies and NGS pipelines. If the SNVs at excess sites found in a particular cancer are due to hypermutable sites then we would expect them to be randomly distributed across research groups (i.e. all research groups should identify the same hypermutable sites). If however the SNVs at excess sites are due to error then we might expect them to be heterogeneously distributed across research groups (i.e. the calling of recurrent false positive SNVs should be systematic of individual research group NGS pipelines). The liver cancers, which were all virus associated hepatocellular carcinomas, , were sequenced by two different groups; 66 from the RIKEN group using the Illumina Genome Analyser (https://dcc.icgc.org/projects/LIRI-JP) and 22 from the National Cancer Centre in Japan using the IIlumina HiSeq platform (https://dcc.icgc.org/projects/LINC-JP). We found that the excess SNVs were heterogeneously distributed amongst research groups (Fisher's exact test, $P = 4x10^{-6}$) suggesting that the 30 excess sites from liver cancers were predominantly errors (Appendix 3.1).

3.4.4 Parameter estimation

To gauge how much variation there is in the density of SNVs across the genome we fit two models to the data using maximum likelihood. In model 1 we allowed the density of SNVs to vary between sites according to a gamma distribution, estimating the shape parameter, and hence the amount of

variation there was between sites. We fitted two versions of this model. In the first version, 1a, we constrained the model such that the mean SNV density, shape parameter, and hence the level of variation, was the same for all triplets. In the second version, 1b, we allowed the mean SNV density and shape parameter to vary between triplets. The second of these models fits the data significantly better than the first according to a likelihood ratio test suggesting that the level of variation differs between triplets (Table 3.2). However, a goodness of fit test, comparing the number of sites predicted to have 1, 2, 3…etc SNVs per site to the observed data, suggests the model fits the data poorly. We therefore fit a second pair of models in which we allowed the rate of SNVs to be due to two processes. The first process, is constant across sites whereas the second process is variable and drawn from a gamma distribution. There are two parameters in the model, the proportion of SNVs at a site produced by the first process and the level of variation in the second process. This model might represent a situation where the rate of mutation is constant across sites but the rate of sequencing error is variable. As with the first model we fit two versions of this model; in Model 2a we constrained the model such that the parameters of the two processes were the same for all triplets. In Model 2b they were allowed to vary between triplets. Both models 2a and 2b fit the data significantly better than models 1a and 1b, and of this second pair of models, model 2b, which allows the parameters to vary between triplets fits the data significantly better than model 2a, in which the parameters are shared across triplets (Table 3.2). The best fitting model is therefore one in which we have two processes contributing to the production of SNVs, one that is constant across sites, although it differs between triplets, and one which is variable across sites. Although, we can formally reject this model using a goodness-of-fit test (Chi-square $p < 0.0001$), because we have so much data, it is clear that the model fits the data fairly well (Figure 3.2). Under this model we estimate that approximately 4.1%, 2.8% and 4.2% of SNVs are due to the process that varies across sites in the TE and NTE, and EX sequences respectively.

Model	N	Log-likelihood	Shape	Median ε
1a	2	-269283	0.13	
1b	64	-2936	0.12	
2a	3	-266889	0.00021	0.956
2b	*96*	*-1302*	*0.00016*	*0.959*

Non-Exon Non-TE (NTE)

Model	N	Log-likelihood	Shape	Median ε
1a	2	-227728	0.31	
1b	64	-1207	0.37	
2a	3	-227026	0.00039	0.963
2b	*96*	*-566*	*0.00026*	*0.972*

Exon (EX)

Model	N	Log-likelihood	Shape	Median ε
1a	2	-13878	0.18	
1b	64	-270	0.22	
2a	3	-13842	0.00019	0.966
2b	*96*	*-240*	*0.00035*	*0.958*

Table 3.2. The fit of 4 models to the observed distribution of recurrent SNVs in the three different genomic fractions A) TE, B) NTE and C) EX. N = number of parameters. Italics indicate the best fit as determined by a liklihood ratio test.

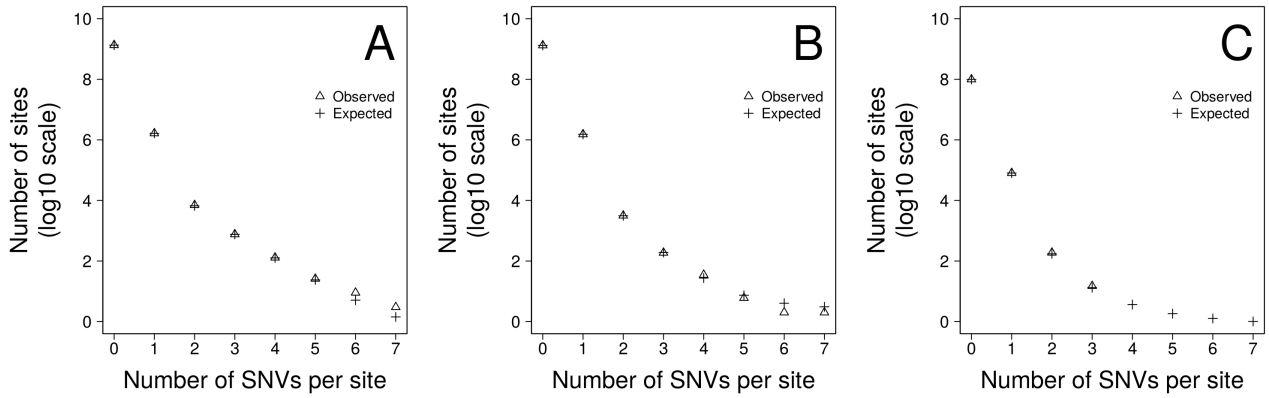


Figure 3.2. The fit of the observed recurrent SNV distribution to expected distribution under the favoured model, 2b, for A) TE, B) NTE and C) EX genomic fractions.

However, the variation in the density between sites due to the variable process is extremely large. The median shape parameters are 0.00016, 0.00026 and 0.00035 for the TE and NTE, and EX sequences respectively. Under a gamma distribution with a shape parameter of 0.0004 we would expect more than 99% of sites to have no SNVs generated by this variable process, but some sites to have a density of SNVs that is 30,000-fold above the average rate.

Discussion.

Through our analysis of ~3 million SNVs from whole cancer genomes we have shown that there are many sites at which there is a significant excess of SNVs. The majority of these are unlikely to be drivers because the density of sites with an excess of SNVs is greater in the non-coding part of the genome than in the exons. It therefore seems likely that the majority of the excess sites are either due to hypermutation or problems with sequencing or the processing of the sequences. Several lines of evidence point to sequencing problems being the chief culprit. First, many of the excess sites disappear when regions of the genome with low mappability are removed. Second, SNVs at a particular excess site tend to be found within the sequences from a particular laboratory; for example, site 85,091,895 on chromosome 5 has 5 SNVs in liver cancers, but all of these are found in the sequences from RIKEN not the sequences from the NCC. It is possbile that this could be caused by biological differences between the cohorts, either environmentally induced or endogenous genetic variation, such as that seen between European and African populations and the differing frequency of 5’-TCC-3’ > 5’-TTC-3' mutation (Harris, 2015). However the level of site and cohort specific, but cryptic, variation required would be huge and we have very little evidence to support such a hypothesis. Third, the level of variation in the density of SNVs is much greater than has been observed or suggested for variation in the mutation rate (Hodgkinson & Eyre-Walker, 2011; Kong *et al.*, 2012; Michaelson *et al.*, 2012) though see a recent analysis of de novo germ-line

mutations which suggests there could be extreme mutational heterogeneity (Smith *et al.*, 2016); some sites are estimated to have rates of SNV production that are tens of thousands of times faster than the genomic average.

Only one line of evidence suggests that there might also be substantial variation in the mutation rate as well as variation in the error rate. When we eliminate sites that are not uniquely mappable at 20bp we find a great reduction in the number of excess sites relative to the case when we remove sites that are not uniquely mappable at 100bp, and yet the read length is greater than 100bp in the data that we have used. This might suggest that there are some repetitive sequences that are prone to a process of hyper-mutation. However, it might also be that mappability at 100bp is not a good guide to mappability during sequence processing. First, some level of mismatch must be allowed during the mapping of reads to the reference because there are single nucleotide variants segregating in the population and there are somatic mutations in cancer genomes. Second, the mappability score is assigned to the first nucleotide of the *k*-mer that can be mapped. Third, although the read length was greater than 100bp, some shorter reads may have been used. Next generation sequencing involves a number of biological processes, such as the polymerase chain reactions in the pre-sequencing creation of libraries and the polymerization of nucleotides during sequencing by synthesis, any one of which can result in technology-specific sequencing artefacts (Quail *et al.*, 2008; Nazarian *et al.*, 2010), in addition to the considerable post-sequencing processing, such as filtering and mapping, which can also generate errors (Harismendy & Frazer, 2009; Minoche, Dohm & Himmelbauer, 2011). Unfortunately it is not possible to say which of these factors is most important.

We have fit two models to the data in which the density of SNVs varies across sites. In the first we imagine that the variation is due to a single variable process and in the second we imagine it is due

to two processes, one of which is constant across sites and one which is variable. We find that this second model fits the data much better than the first model, although it can be formally rejected by a goodness-of-fit test. In this second model we estimate the proportion of SNVs that are due to the two processes and the level of variation. We estimate that approximately 2.8-4.2% of SNVs are due to the second process and that this second process is highly variable between sites, such that a few sites have a density of SNVs that is ten of thousands higher than the average density. It is possible that the first process is mutation and the second is sequencing error, but we cannot rule out the possibility that the second process includes variation in the mutation rate as well. Studies of germ-line (Hodgkinson & Eyre-Walker, 2011; Michaelson *et al.*, 2012) and somatic (Hodgkinson, Chen & Eyre-Walker, 2012; Woo & Li, 2012; Lawrence *et al.*, 2013; Liu, De & Michor, 2013; Polak *et al.*, 2015) mutations have indicated that the mutation rate varies between sites on a number of different scales. However, indications are that the variation is probably fairly modest (Hodgkinson, Chen & Eyre-Walker, 2012; Michaelson *et al.*, 2012).

A model including two processes fits the data well (figure 2). However, e can reject this model in a goodness-of-fit test, because we have a huge amount of data. Possible reasons for the less than perfect fit include large scale variation in the mutation rate (Hodgkinson & Eyre-Walker, 2011; Schuster-Bockler & Lehner, 2012; Makova & Hardison, 2015) and multi-nucleotide-mutations (MNMs) (Rosenfeld, Malhotra & Lencz, 2010; Schrider, Hourmozdi & Hahn, 2011; Harris & Nielsen, 2014); the latter represent ~2% of all human single nucleotide polymorphisms (SNPs)
In conclusion it seems likely that many sites in somatic tissues that have experienced recurrent SNVs are due to sequencing errors or artefacts of post-sequencing processing and there seems to be little evidence of cryptic variation in the somatic mutation rate. However, this not necessarily mean that such variation does not exist – it would be extremely difficult to detect it given the high level of site-specific sequencing error. As sequencing technology and processing pipelines improve in

accuracy, we would expect similar future analyses to be able to confidently estimate the true underlying variation in the somatic mutation rate. Accompanied by the flow of data from projects such as the 100k genomes project, it should soon be possible to achieve per triplet mutation rate variation map for individual cancer types and not just pooled across multiple cancers.

4. Mutation rate variation is not associated with common fragile site expression.

4.1 Abstract

Aphidiclolin-induced common fragile sites (aCFS) are large genomic regions that recurrently appear as breaks on metaphase spreads. They have been shown to coincide with regions of large-scale genomic instability that are susceptible to copy number variants and inter-chromosomal rearrangements. Small mutations (SM), such as point mutations and small insertions and deletions are also ubiquitous in cancer genomes, representing the genomic instability and mutational processes within the cancer. The occurrence of both of these processes have been linked to replication stress; One consequence of replication stress is replication fork breakage followed by attempts to restart replication using the homologous recombination machinery, which can remain error prone after restart. We hypothesised that the genomic instability at aCFSs in cancers could result in broken replication forks, followed by error prone replication restart and thus an increase in SMs at these loci. To investigate this we analysed the distribution of >3 million SMs from 507 whole cancer genomes in aCFSs and non-fragile regions. We found no strong evidence to indicate that aCFSs are enriched in SMs beyond the expected influence of replication time, and thus conclude that the biological processes contributing to these different forms of large-scale and small-scale mutation, are unlikely to be linked and may suggest that, while replication fork breakage is a feature of loci prone to fragility, replication fork restart may besuppressed at these loci.

4.2 Introduction

Fragile sites (FS) are unstable, mega-base-long genomic regions that recurrently appear as breaks on metaphase spreads. The first fragile site described was found on chromosome 9 in blood cultures in 1965 (Dekaban 1965). In 1984, aphidicolin, an agent of replication stress, was shown to induce repeated chromosomal breaks at specific locations in lymphocytes, and the term common fragile site (CFS) was coined (Glover *et al.* 1984). Recent work, mostly carried out in lymphocytes, has uncovered many features associated with CFSs (reviewed in (Debatisse *et al.* 2012; Ozeri-Galai *et al.* 2012; Durkin & Glover 2007)). They are reported to be associated with high AT content, to occur in late replicating regions (Letessier *et al.* 2011; Le Tallec *et al.* 2011) and to co-localise with large genes or transcription units (Le Tallec *et al.* 2013; Smith *et al.* 2006; Wilson *et al.* 2015).

Loci associated with CFSs are also prominent in cancer etiology. They have been implicated in the reduced expression of overlapping tumour suppressor genes (Smith *et al.* 2006), are associated with large-scale mutagenic events in tumours, including large insertions and deletions and inter-chromosomal rearrangements, and are prone to display copy number variations (CNVs)(Beroukhim *et al.* 2010; Bignell *et al.* 2010; Gao *et al.* 2014). Such genomic rearrangements can be broadly classified into those that result from non-allelic homologous recombination between relatively extensive (>100bp) regions of sequence identity and those that display no homology or limited microhomology at the their breakpoint junctions (Liu *et al.* 2012). The formation of CNVs, both in the germline (Lee *et al.* 2007) and in somatic tissues such as cancer (Arlt *et al.* 2011) has been linked to replication problems.

Single nucleotide variants (SNVs) or small insertions/deletions (<100 base pair indels) – collectively referred to hereon as small mutations (SM) - are also ubiquitous in cancer genomes.

Most SM are considered to be “passengers” or “noise” and represent unselected mutational processes of unstable genomes (Alexandrov *et al.* 2013; Lawrence *et al.* 2013; Stratton *et al.* 2009). The rate at which SNVs arise varies massively between cancers (Lawrence *et al.* 2013), and across the genome the variation in SNV density is known to correlate with GC content, replication time, distance to telomeres and histone modifications (Hodgkinson *et al.* 2012; Schuster-Bockler & Lehner 2012). However, these established correlations are only able to account for 40-60% of the variation observed (Hodgkinson *et al.* 2012; Hodgkinson & Eyre-Walker 2011; Schuster-Bockler & Lehner 2012) with replication time reported as the dominant correlate (Liu *et al.* 2013; Woo & Li 2012; Koren *et al.* 2012). Although indels have been understudied in this area, likely due to technological difficulties in detection (Jiang *et al.* 2015; Mullaney *et al.* 2010), germ-line studies have shown that their density along the genome varies in a similar fashion to SNVs, with hot-spots exhibiting elevated indel rates up to 24-fold greater than the regional average (Mills *et al.* 2006) and contribute to a similar level of genetic variation as SNVs (Mullaney *et al.* 2010; Mills *et al.* 2011).

Most of the models that have been proposed to explain the origin of CFSs and large chromosomal rearrangements propose that replication has been compromised as a result of either DNA secondary structure (Walsh *et al.* 2013) or clashes between replication and transcription (Wilson *et al.* 2015). Because there is a paucity of origins associated with CFSs, this has been proposed to result in replication-associated DNA structures that are prone to recombination. Recombination mediated replication restart events in model organisms have been demonstrated to result in both non-allelic homologous recombination (Lambert *et al.* 2005; Lambert *et al.* 2010) and, even when restart occurs at the correct locus, in error prone replication (Deem *et al.* 2011; Sakofsky *et al.* 2015; Mizuno *et al.* 2012). Errors resulting from homologous recombination-restarted replication include template switching events, replication slippage at microhomology (Deem *et al.* 2011) and a variety of other SMs. Such errors have been proposed to underpin the etiology of a number of human

genetic disorders (Hastings *et al.* 2009). Currently, there have been no reports of an association between CFSs and SMs in cancer cells, despite the fact that both CFSs and SMs are associated with genomic instability and replication time in cancers. We hypothesize loci associated with aphidicolin induced CFSs (aCFS) would be enriched in SMs in cancer samples. To this end, we examined the density of SMs from 507 human cancer whole genomes in aCFS and non-FS regions.

4.3 Methods

4.3.1 Replication timing data and dividing the genome into 100 kilo-base windows

We downloaded Encode Repli-seq wavelet smoothed signal data (Thurman *et al.* 2007; Hansen *et al.* 2010) aligned to the reference genome (HG19/GRCh37) for the GM12878, HeLa, HUVEC, K562 and HepG2 cell lines from the UCSC ftp site ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ to represent a wide range of cell types. The GM12878 cell line is an Epstein-Barr virus infected B-lymphocyte cell derived from female blood tissue with a normal karyotype, the HeLa cell line is from a cervical carcinoma, the HUVEC cell line is derived from human umbilical vein endothelial cells with a normal karyotype, the K562 cells were established from the blood tissue of a female with leukaemia, and the HepG2 cells are derived from the liver tissue of a male with hepatocellular carcinoma (http://genome.ucsc.edu/encode/cellTypes.html). We computed the mean replication time for chromosomes 1-22 and X in non-overlapping 100 kilo-base windows across 5 cell lines (GM12878, HeLa, HUVEC, K562, HepG2) for correlations with SNV densities. We used the Repli-seq start coordinates for each chromosome (bp 24500) to denote the beginning of the first window. Chromosome Y was not included as no replication time data was available. Any windows

containing unknown bases (denoted as N in the reference genome) were excluded from the analysis. This resulted in 28,080 100 kilo-base windows.

4.3.2 Mutation densities

SNVs were taken from publicly available somatic mutations called from 507 whole genome sequences (Alexandrov *et al.* 2013) at ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/mutational_catalogs/genomes. The initial 3,382,737 SNVs were filtered to exclude those for which no replication time data existed, including those on the Y chromosome to produce a final count of 3,348,815 SNVs used in the analysis. 214,058 indels were taken from 265 whole genome sequences from the same source as the SNVs and subjected to the same filters, resulting in a total of 210,314 indels used in the analysis. Indel and SNV densities for each 100 kilo-base window were calculated as the total number of mutations in that window divided by the number of genomes (507 for the SNVs and 265 for the indels). No correction for unannotated bases was required as any windows with an unannotated base were removed from the analysis.

4.3.3 Compiling fragile sites

Four fragile site datasets were collated from i) Mrasek *et al.* (2010) (Mrasek *et al.* 2010) (henceforth referred to as MRA), ii) Lukusa & Fryns (2008) (Lukusa & Fryns 2008) (henceforth referred to as LUK), iii) Le Tallec *et al.* (2013) (Le Tallec *et al.* 2013), (henceforth referred to as LET) and iv) Bignell *et al.* (2010) (Bignell *et al.* 2010) (henceforth referred to as BIG). Genomic coordinates for BIG were taken from the refined mapping position in their supplemental data and converted to HG19 from HG18 using the lift over utility available from University of California, Santa Cruz at http://genome.ucsc.edu/ (Hinrichs *et al.* 2006). LET, LUK and MRA HG19 genomic coordinates

were obtained directly from cytoband information listed in each respective publication. Only BIG, LUK and LET were considered for the aphidicolin-induced fragile site (aCFS) analysis as most MRA FSs occur at a low frequency. The inverse of the MRA coordinates were used in the classification of non-FS regions. Due to the low resolution to which FS are mapped, each 100 kilo-base window was only classified as being contained within a FS if the whole window resided inside the FS coordinates. Non-FS windows were those that were non-FS in all four datasets.

4.3.4 Statistical analysis.

The R stats package, R version 3.3.0, was used for all statistics (R Core Team 2016). To compare SNV and indel densities between various test and control data, median densities were used and significance was tested using the Mann-Whitney test. Differences in replication time between datasets were also tested using the Mann-Whitney test. Spearman's rank correlation was used to test for associations between replication time and SNV density for the each of the K562 aCFSs. Linear and Quadratic regressions were modelled using the “lm” function with replication time as the predictor and SNV density as the response, the best fitting model was selected by the lowest AIC value. A t-test was used to test if the SNV densities in the latest replicating deciles of the non-FS control and aCFS tests were significantly similar.

4.4 Results

4.4.1 Classifying fragile sites

In trying to assess the density of SM at loci that are representative of fragile sites (FS), the first challenge is to define what genomic regions are considered to be FS. Investigations in cells other than lymphocytes (Le Tallec *et al.* 2011; Le Tallec *et al.* 2013) and experiments with other inducers of FSs, such as 5-Azacytidine and BrdU (See (Baskaran & Brahmachari 2000) for a review) produce varying classifications of FSs. For example, data from one study (Mrasek *et al.* 2010), which includes infrequently expressed aphidicolin induced FSs, would estimate that ~66% of the genome is fragile. We therefore consider only aphidicolin induced common fragile sites (aCFS) with a break frequency of >1% for further analysis. However, even within this definition, multiple studies have produced different lists of aCFSs. We therefore investigate three possible aCFS datasets, each with their own merits; i) the LUK dataset (Lukusa & Fryns 2008) which compiled the largest collection of 76 aCFSs in lymphocytes, covering ~16% of the genome, ii) the BIG dataset (Bignell *et al.* 2010) which compiled data on high resolution FISH mapping of 38 aCFSs covering ~7% of the genome, and iii) the LET dataset (Le Tallec *et al.* 2013) which compiled 50 aCFSs across 9 different tissue types covering ~12% of the human genome. Each of the data sets; LUK, BIG and LET, contain aCFSs that are not only private to that data set but also shared across datasets, therefore direct comparison between datasets is not simple. Additionally the higher resolution mapping of the BIG data means that it is not directly comparable to the LUK and the LET data. We can however compare these 3 datasets individually to genomic regions considered to be non-fragile (non-FS). We considered non-FSs to be those genomic regions that were not an aCFS in any of the LUK, BIG or LET datasets nor ever reported to be one of the 230 FSs as reported by Mrasek *et al.* (Mrasek *et al.* 2010).

4.4.2 SM density in fragile sites

We divided the genome into non-overlapping 100 kilo-base windows and scored each window as

being wholly contained within either a LUK (4940 windows), BIG (2073 windows) or LET (3931 windows) aCFS or non-FS (8581 windows) if the window was not an aCFS in any of the four datasets. Windows that were not defined as either, i.e. reported to harbour infrequently expressed FSs by Mrasek *et al.* (Mrasek *et al.* 2010), totalling 41% of the genome, were excluded from the analysis. The density of SNVs and Indels were then calculated for each window. Figure 4.1 shows the median density of SNVs and indels for each dataset. We found that the median SNV density for LUK (0.183 SNVs/100kb) and BIG (0.189 SNVs/100kb) were significantly lower, but did not change dramatically, from the non-FS controls (0.215 SNVs/100kb) (Mann-Whitney Test, $p < 0.001$), whereas the LET SNV density was significantly enriched by 25% (0.268 SNVs/100kb) compared to the non-FSs (Mann-Whitney Test, $p < 0.001$). In no dataset was median indel density found to be different from the control (0.0264 indels/100kb).

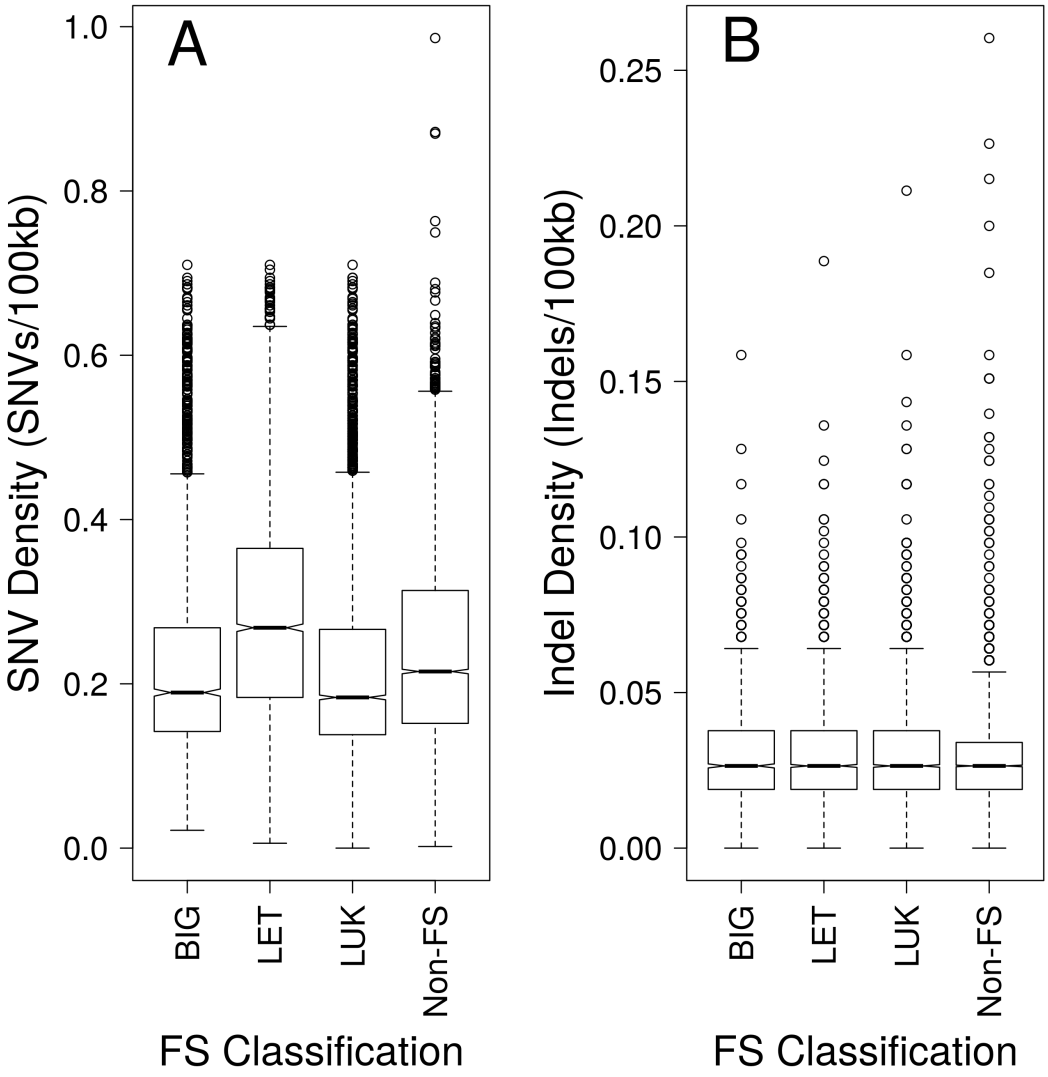


Figure 4.1. Boxplots comparing A) median SNV density and B) median indel density for each dataset and the non-FS data.

We examined the LET aCFSs further to try and determine the origin of this SNV enrichment. The LET data combines aCFSs from 9 different tissue types so we first investigated if the enrichment came from aCFSs identified in any particular tissue. SNV density in aCFSs from 7 of the 9 tissue types was significantly enriched from the non-FS control (Mann-Whitney test, $p < 0.001$) (Figure 4.2), however many of these differences were modest and the variation around the medians were large. The most significant difference (57% enrichment) was in aCFSs from K562 cells. Interestingly, the SNV density in lymphocytes, the cells used for the aCFSs in LUK and the cells in which most studies have been carried out, were significantly depleted compared to the non-FS control. Only the SNV density in aCFSs from LS174T cells, a colonic adenocarcinoma cell line, was not significantly different from the non-FS control.

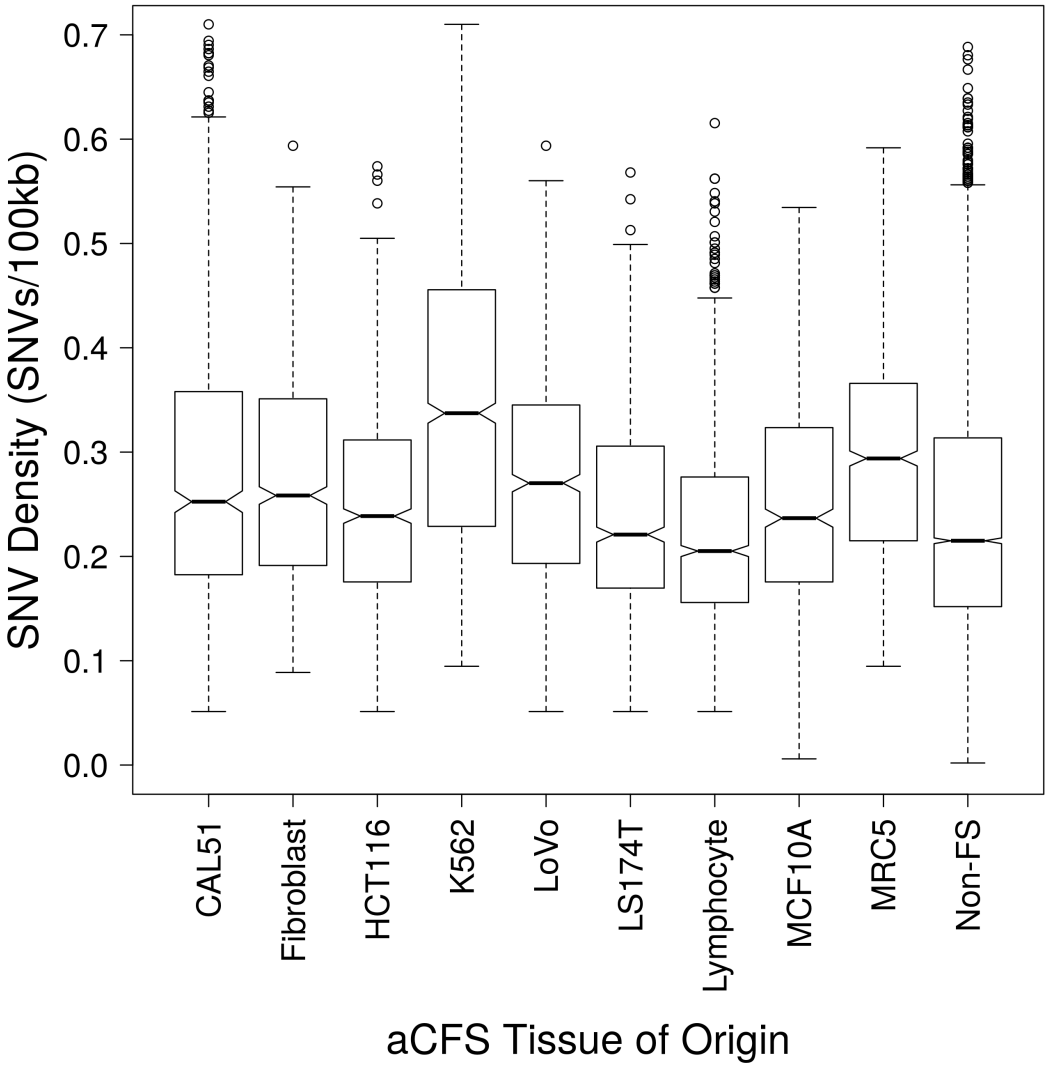


Figure 4.2. Boxplot showing variation in SNV density in aCFSs from each of the 9 tissues types from comprising the LET data and the non-FS data (far right).

4.4.3 SNV density and replication time

It has been shown that replication time correlates strongly with SNV density in somatic tissues (Liu *et al.* 2013; Woo & Li 2012; Schuster-Bockler & Lehner 2012; Koren *et al.* 2012) and that it is important in aCFS etiology (Durkin & Glover 2007; Letessier *et al.* 2011; Le Tallec *et al.* 2011). Due to the high enrichment of SNVs in aCFSs from the K562 cell line, we examined the relationship of SNV density with replication time for each of the K562 aCFSs. The median replication times of each of the 13 K562 aCFSs showed the majority of K562 aCFSs are late replicating and that this is highly negatively correlated with the median SNV density (Spearman's correlation coefficient $r = -0.87$, $p < 0.0001$) (Figure 3). Interestingly the ~10 mega-bases covering FRA16D, the only aCFS found in every tissue type, has a median replication time and SNV density very close to that of the genomic average.

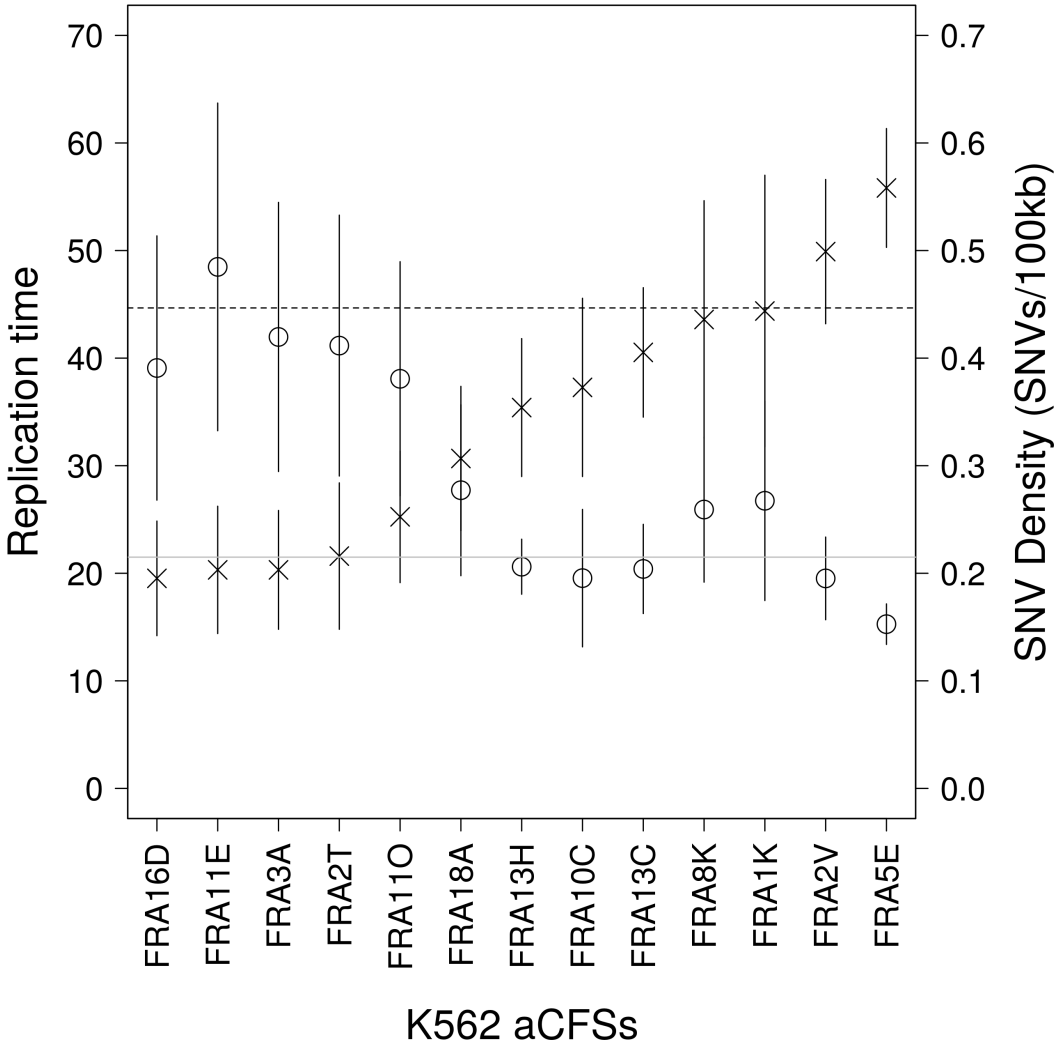


Figure 4.3. Association between median replication time (circles) and median SNV density (crosses) for each of the 13 aCFSs found in K562 cells ranked by SNV density. Vertical lines represent one median absolute deviation. Horizontal lines are the non-FS median replication times (dashed) and SNV density (solid).

We thought the density of SNVs in the K562 aCFSs may be simply a consequence of replication time. To address this we ran regression models to factor out the influence of replication time on SNV density. Using data from all 28,080 windows, combining both aCFS and non-FS windows. We found that a regression model involving a quadratic term fit the data significantly better than a model with only a linear term (ΔAIC = 3016) (Figure 4.4a). However, we also found that a model in which the K562 aCFS and non-FS had their own relationship between SNV density and replication time fit the data better (ΔAIC = 391) (Figure 4.4b), suggesting that the relationship between SNV density and replication time differs significantly between the two categories of sites. This difference seems to be largely down to an enrichment in SNVs in late replicating regions of the genome in aCFS.

It is apparent that the greatest enrichment is in late replicating DNA (Figure 4.4b), we find that in the latest replicating decile, SNV density is enriched 19% over the effect expected from replication time alone (Table 4.1). If this enrichment is a canonical feature of aCFSs then we would expect to see this pattern replicated for the other LET tissue types and for LUK and BIG. This is not the case. We find considerable variation in ΔAICs when comparing the models for cell type/dataset compared to the non-FS control, and in the enrichment of SNV density in the latest replicating decile, where the ratios of SNV density range from 0.76 – 1.19 (Table 4.1, Appendix 4.1). Fibroblasts are the only tissue type in which the distribution of SNV density is not significantly different to the non-FS control (two-sample t-test) (Table 4.1). Thus we show that variation in SNV density across aCFSs can not be explained fully in terms of replication time and that increased SNV density is not a canonical feature of aCFSs.

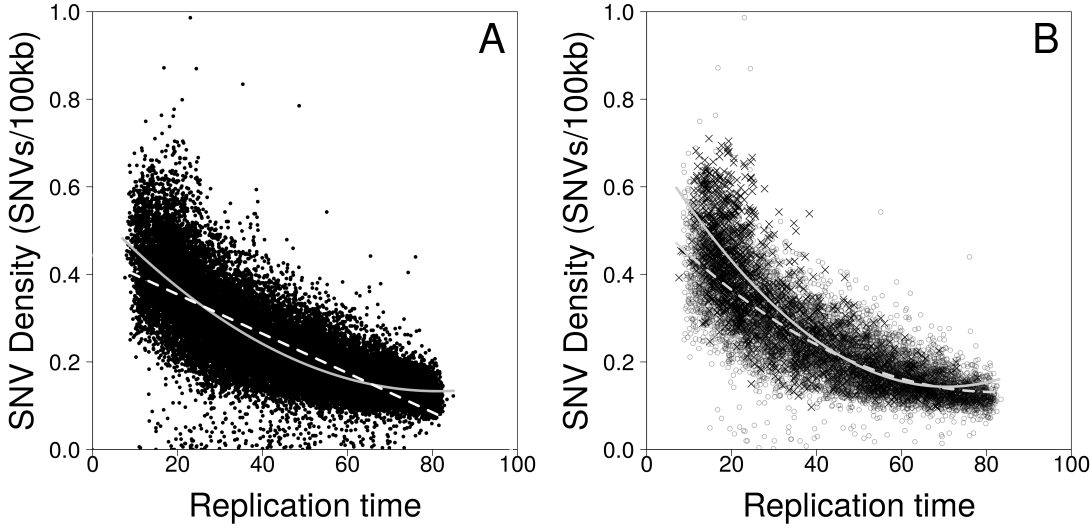


Figure 4.4. A) The fit of linear (dashed line) and quadratic (solid line) models to mean SNV density as a product of replication time for all 28,080 100 kilo-base windows. Each point represents one window. B) Right - The fit of separate quadratic terms for K562 aCFS (solid line and crosses) and non-FS (dashed line and circles) windows.

Tissue Type / Datset	QM1 AIC	QM2 AIC	ΔAIC	aCFS SNV Density	Non-FS SNV Density	Decile Enrichment	T-Test t(1)	p value
K562FS	-24826	-25217	-391	0.50	0.42	1.19	-11.80	<0.0001
BIG	-27542	-27568	-27	0.48	0.41	1.15	-6.13	<0.0001
CAL51	-23801	-23804	-3	0.48	0.42	1.15	-4.97	<0.0001
LUK	-35644	-35660	-16	0.44	0.41	1.07	-4.10	<0.0001
Fibro	-24932	-24939	-8	0.42	0.41	1.01	-0.34	0.738
MRC5	-25184	-25184	0	0.38	0.42	0.91	4.92	<0.0001
LoVo	-24495	-24534	-39	0.37	0.42	0.88	6.41	<0.0001
HCT116	-25012	-25128	-116	0.34	0.41	0.81	9.43	<0.0001
MCF10A	-24125	-24310	-185	0.34	0.42	0.81	7.86	<0.0001
LS174T	-24800	-24928	-128	0.33	0.42	0.80	12.07	<0.0001
Lymph	-25916	-26018	-102	0.31	0.41	0.76	14.28	<0.0001

Table 4.1. Comparison of fitting a quadratic model to variation in SNV densities in aCFS from different tissue types and datasets. QM1 AIC = AIC values for a quadratic model for aCFS and Non-FS combined, QM2 AIC = AIC values when allowing a separate model for aCFS and Non-FS, ΔAIC = QM2 – QM1, aCFS SNV Density = aCFS SNV density in the latest replicating decile of windows, Non-FS SNV Density = non-FS SNV density in the latest replicating decile of windows, p value = p value obtained from a two-sample t-test comparing the latest replicating aCFS and non-FS deciles.

4.5 Discussion

Fragile sites (FS) have been shown to coincide with regions of large-scale genomic instability that are susceptible to CNVs and inter-chromosomal rearrangements (Beroukhim *et al.* 2010; Bignell *et al.* 2010). FS expression and large-scale genomic rearrangements are both linked to replication stress. One consequence of replication stress is replication fork breakage followed by attempts to restart replication using the homologous recombination machinery (Lambert *et al.* 2005; Petermann *et al.* 2010). Errors during this process are known to promote non-allelic recombination, one of the causes of large-scale rearrangements (Lambert *et al.* 2005; Lambert *et al.* 2010; Mizuno *et al.* 2009). In several model organisms, once restarted, the replication machinery remains error prone, being subject to microhomology mediated template switching, replication slippage and the production of elevated levels of SMs (Deem *et al.* 2011; Sakofsky *et al.* 2015; Mizuno *et al.* 2012; Iraqui *et al.* 2012). Our hypothesis was that, if FSs and large-scale genomic rearrangements have an etiology in replication fork collapse and restart, loci prone to these phenomena would also display an increase in the small mutations that are characteristic of the restarted replication. We thus investigated if loci associated with aCFSs and large-scale chromosome rearrangements in cancer cells also displayed increased SMs. We found no convincing evidence that this is the case, for either indels or SNVs. This may highlight a distinction between the biological pathways involved in the etiology of small-scale and large-scale mutagenic events and may suggest that, while replication fork breakage is a feature of loci prone to fragility, replication fork restart is suppressed at these loci.

Defining which regions of the genome are fragile for such an analysis is not straightforward. Multiple studies across different tissue types define different regions as being FSs. We simplified the matter somewhat by considering only aCFSs for this analysis, however this still left us with data

from 3 separate studies, which differed in which regions of the genome were considered as aCFSs. None of the datasets showed an increase in indel density within aCFSs, although this could be due to technical difficulties in detection of indels from next generation sequencing data (Jiang *et al.* 2015), and data from two of the studies, BIG and LUK, showed no enrichment of SNVs in aCFSs, however initial results from LET suggested that SNVs may be enriched in aCFSs. This was interesting as it was the only study to date to investigate the occurrence of aCFSs in tissues other than fibroblasts or lymphocytes, covering 9 different cell types in total. Further investigation showed that the increase was coming from only a subset of these cell types, with the highest enrichment found in K562 cells. When examining the 13 aCFSs found in K562 cells, it was apparent that this increase in SNV density was likely due to the late replicating nature of the K562 aCFSs. This was largely confirmed when regressing SNV density as a function of replication time, however we also detected variation in SNV density within aCFSs that was not attributable to replication time, although it was most prominent in later replicating regions. Some of this variation could be due to our methods of using the average replication times across 5 cell lines; multiple cell types have been shown to exhibit different replication time profiles (Morganella *et al.* 2016) and disruption to these profiles throughout differentiation (Lu *et al.* 2014), however the extreme replication time deciles exhibit remarkable conservation and stability across cell types (Morganella *et al.* 2016). Therefore to minimise any potential effect from differing replication time profiles, we considered the enrichment of SNVs only in the latest replicating deciles. We found that the K562 aCFSs were moderately enriched in SNVs, but that this pattern was not common to all tissue types or data sets, and therefore does not appear to be a canonical feature of an aCFS.

We also note that the one ubiquitous aCFS, FRA16D, has an average replication time and SNV Density close to to the control. At first glance this would seem to suggest that, contrary to previous findings (Durkin & Glover 2007; Ozeri-Galai *et al.* 2012; Letessier *et al.* 2011), late replication is

not an indicator of fragility. This however highlights the importance of considering scale in an analysis. It has been previously demonstrated (Le Tallec *et al.* 2011) that the effect of late replication on FS expression in FRA16D happens over one or two mega-bases that have a paucity of replication origins and are replicated by forks traversing long distances. Thus the average replication time for an FS region may be inconsequential for FS expression, the fact that it contains a mega-base or two of origin poor, late replicating, DNA may be enough to express the FS if the replication origins flank this region.

The majority of previous studies into aCFSs have been carried out in lymphocytes. This analysis further supports previous assertions (Le Tallec *et al.* 2011) that tissue specificity is important to consider when investigating FSs, some aCFS are private to a particular tissue type whereas many aCFS are shared among tissues but at varying frequencies (Le Tallec *et al.* 2011; Le Tallec *et al.* 2013). This likely represents the specific biology of individual cell types, demonstrated by the differing replication profiles of FRA3B in Fibroblasts and Lymphocytes (Le Tallec *et al.* 2013).

This analysis uses SMs from a variety of cancer samples, many unrelated to the tissues in which these aCFSs were discovered, Therefore the tissue specificity of aCFS expression could be obscuring any associations with SMs, however we think this is unlikely to be the case for three reasons. First, a large number of aCFS are shared between tissue types, so if there was any association between SM and aCFS expression we would expect to see at least a small indication of this in the shared aCFS such as FRA3B and FRA16D, however we do not. Second, replication time is known to be strongly correlated with SNV density, so by focusing on the latest replication time decile, which is largely invariant across tissue types (Morganella *et al.* 2016), we should minimise any tissue specific effects. Third, previous studies have found associations of other features, such as an overlap of homozygous deletion clusters with aCFSs, when not considering tissue specificity

(Bignell *et al.* 2010). This final point suggests that although tissue specific FS exist, many determinants of fragility may not be tissue specific.

Here we have shown that aCFSs are not specifically enriched in either indels or SNVs in cancers and that any apparent enrichment of SNVs in aCFS is largely explained by replication time, and other, unknown components of moderate effect size. Our analysis thus does not support a hypothesis that genetic instability associated with FSs in cancer cells is significantly associated with attempts to restart broken forks by homologous recombination. We also show that when considering replication time as a correlate of any phenomenon, resolution is an important aspect. Average replication times across multiple mega-bases will capture much of the correlation with SNVs but will obscure important features that are pertinent to biological events, such as FS expression at a finer scale.

5. The relationship between divergence, diversity and the rate of *de novo* mutation along the human genome.

5.1 Abstract

The rate of divergence between species, and the level of diversity within a species, are expected to depend on the rate of mutation. Here we investigate the relationship between divergence, diversity and the rate of mutation across the human genome at scales of 1Mb and 100kb using >40,000 *de novo* mutations (DNM). We show that there is significant variation in the rate of DNM across the human genome, but the variation is modest. We confirm that the rate of divergence along the human lineage is correlated to the rate of DNM but that the correlations are weaker than expected. We provide evidence that this is due the effect of biased gene conversion (BGC) on the probability that a mutation will become fixed. In contrast we find that diversity is almost as strongly correlated to the rate of DNM as it can be given the sampling error in the number of DNMs. It therefore seems that variation in the mutation rate is the chief determinant of the level of diversity across the human genome at the 1Mb and 100kb scales, and this suggests, contrary to previous analyses, that recombination does not affect diversity in humans independently of its effect on the mutation rate. Finally we show that the correlation between divergence and DNM density declines as increasingly divergent species are considered. Our results have important implications for the use of divergence and diversity data to study variation in the mutation rate.

5.2 Introduction

Until recently, the distribution of germ-line mutations across genome was largely studied using patterns of nucleotide substitution between species in putatively neutral sequences - see (Hodgkinson & Eyre-Walker 2011 for review of this literature), since under neutrality the rate of substitution should be equal to the mutation rate. However, the sequencing of hundreds of individuals and their parents has led to the discovery of thousands of *de novo* mutations (DNMs) in humans; it is therefore possible to start analysing the pattern of DNMs themselves rather than inferring their patterns from substitutions. Initial analyses have shown that the rate of DNM increases with paternal age (Neale *et al.* 2012; O'Roak *et al.* 2012; O'Roak *et al.* 2011; Iossifov *et al.* 2014; Iossifov *et al.* 2012; Michaelson *et al.* 2012; Kong *et al.* 2012; Francioli *et al.* 2015; Wong *et al.* 2016), a result that was inferred by Haldane some 70 years ago (Haldane 1947), that it varies across the genome (Michaelson *et al.* 2012) and that it is correlated to a number of factors, including the time of replication (Francioli *et al.* 2015) the rate of recombination (Francioli *et al.* 2015), GC content (Michaelson *et al.* 2012) and nucleosome occupancy (Michaelson *et al.* 2012).

Although, divergence between species has been used to study the pattern of DNM, it has recently been shown that the rate of mutation, at a per Mb scale, is not as strongly correlated to the rate of nucleotide substitution between species as it could be (Francioli *et al.* 2015). Instead, the rate of substitution appears to be more strongly correlated to the rate of recombination. This might be due to one, or a combination, of several factors. First, recombination might affect the probability that a mutation becomes fixed by the process of BGC (Duret & Galtier 2009). Second, recombination can affect the probability that a mutation will be fixed by natural selection; in regions of high recombination advantageous mutations are more likely to be fixed but slightly deleterious mutations are less likely to be fixed. Third, low recombination leads to high levels of genetic hitch-hiking and

background selection, both of which can reduce the time to coalescence in the human-primate ancestor, and hence the divergence between species. And fourth, the independent correlation of divergence and recombination might simply be due to problems with multiple regression; spurious associations can arise if multiple regression is performed on two correlated independent variables that are not know without error.

The object of the current study is to examine the large scale distribution of DNMs and to address the following questions: (i) to what extent does the mutation rate vary across the human genome at a large scale; (ii) does the variation at these scales vary between mutational types; (iii) why does divergence between species depend upon the rate of recombination independently of the DNM rate, (iv) what is the relationship between human diversity and DNM rate, and (v) does the correlation between the rate of DNM and divergence vary between different pairs of mammalian species. Because many more DNMs have been published recently we are able to investigate these questions at both the 1Mb scale, and a finer scale of 100kb.

5.3 Materials and methods

5.3.1 Alignment comparisions and filtering of alignments.

Three sets of alignments were investigated, all based on human genome build hg19/GRCh37 for autosomes only, (i) the University of California Santa Cruz (UCSC) pairwise (PW) alignments (Chiaromonte *et al.* 2002) for human-chimpanzee hg19-panTro4 downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsPanTro4/ (ii) UCSC MultiZ (MZ) 46-way alignments (Blanchette *et al.* 2004) downloaded from

http://hgdownload.cse.ucsc.edu/goldenpath/hg19/multiz46way/ and (iii) Ensembl Enredo, Pecan, Ortheus (EPO) 6 primate multiple alignment, release 74, (Yates *et al.* 2016; Paten *et al.* 2008) downloaded from ftp://ftp.ensembl.org/pub/release-74/emf/ensembl-compara/epo_6_primate/. For the PW and MZ alignments, divergence as a function of mean alignment length - defined by the number of aligned bases for each alignment row in the alignment source files - was calculated per Mb window. This was not possible for the EPO alignments as they are constructed using global alignment methods as opposed to the local alignment methods for MZ and PW. To investigate the consistency of results produced by all three alignment methods, we correlated divergence versus DNM density, iteratively increasing the filter for the minimum number of interrogable bases per Mb window from 1 to 950,000 in 5000 base-pair steps. Interrogable bases were defined as the number of bases defined as A,C,T or G in both species, excluding any gaps "-" or unknown nucleotides "N".

5.3.2 Filtering of EPO alignments and construction of main data set.

The filters applied to the alignments have the to potential significantly impact results. From the results of the alignment comparison, the EPO alignments provided consistent results over a wide range of minimum interrogable bases, remaining relatively unchanged from 200,000 to 900-000 interrogable bases. We chose a filter of 50% interrogable bases per window as the best compromise to retain the maximum number of windows (2,337 out of an initial 2,922) whilst providing consistent results. We also removed all windows that did not contain any data for recombinaton rate.

5.3.3 Selection and filtering of DNMs.

From four sources we obtained 43,433 Autosomal DNMs, 547 from Michaelson *et al.* (Michaelson

et al. 2012), 4,931 from Kong *et al.* (Kong *et al.* 2012), 11,016 from Francioli *et al.* (Francioli *et al.* 2015) and 26,939 from Wong *et al.* (Wong *et al.* 2016). After filtering against interrogable bases in the EPO alignments and the filtering methods mentioned in section 5.3.2, this was reduced to 35,401 DNMs for the human/chimpanzee/orang-utan comparison, 31,185 DNMs for the human/orang-utan/macaque comparison and 23,534 DNMs for the human/macaque/marmoset comparison.

5.3.4 Selection and filtering of SNPs.

All SNPs from the 1000 genomes project phase 3 (The 1000 Genomes Project Consortium 2015) were downloaded from hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/. After removing all multi-allelic SNPs and, structural variants and indels we were left with 77,818,368 autosomal SNPs. After filtering out windows with no recombination rate scores and <50% interrogable bases per window we were left with 71,917,321 SNPs.

5.3.5 Genomic features.

Male specific standardised recombination rate data (Kong *et al.* 2010) was downloded from http://www.decode.com/additional/male.rmap which provides recombination rates in 10kb steps. For each 100kb and 1Mb window the recombination rate was calculated as the mean of these scores with a score assigned to the window in which the position of its first base resided. The number of scores used to calculate the mean of the window was also recorded and as noted in section 5.3.2 was used in the filtering process. We downloaded Encode Repli-seq wavelet smoothed signal data (Hansen *et al.* 2010; Thurman *et al.* 2007), provided in 1kb steps, for the GM12878, HeLa, HUVEC, K562, MCF-7 and HepG2 cell lines from the UCSC ftp site

ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/. We computed the mean replication time for all autosomes for 100kb and 1Mb windows across all 6 cell lines. Replication time were assigned to windows based upon their start coordinates. GC content was calculated directly from the human genome (hg19/GCRh37) for 100kb and 1Mb windows. Nucleosome occupancy for the GM12878 cell line was used (Encode Project Consortium *et al.* 2012), downloaded from /hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhNsome/. Nucloesome occupancy scores are provided at high, but variable resolution, with scores spanning 1 to 27,362 bases. Mean nucleosome occupancy was calculated per 100kb and 1Mb window, accounting for this variation.

5.3.6 Statistical analysis.

The R stats package, R version 3.3.1, was used for all correlations and regression analyses of observed variables (R Core Team 2016).

For bootstrapping we resampled 2500 windows with replacement to find the slope of the regression of normalised divergence and recombination rate, and normalised DNM frequency and recombination rate using the `lm()` function in R. We then minused the slope of the divergence from the slope of the DNM frequency to fine the differnece in the slopes. We performed 1000 replicates and calculated the p-value as the proportion of replicates for which the difference in slope was negative, or 1 minus the proportion of replicates for mutational types where the slope of divergence and recombination rate is greater than that of DNM frequency and recombination rate. Simulations to derive expected variables and comparisons to observed variables were done using Mathematica (version 7).

To estimate the mutation rate distribution we use the method of [43]. In brief we assume that the mutation rate in each block is αu where u is the average mutation rate per site and α is the rate above or below this mean. α is assumed to be gamma distributed. The number of mutations per block is assumed to be Poisson distributed with a mean $\alpha u l$ where l is the length of the block. This means that the number of mutations per block is a negative binomial. We fit the distribution using maximum likelihood using the `NMaximize` function in Mathematica (version 7).

We investigated the correlation between different types of mutation across blocks by fitting a single distribution to both types of mutation; i.e. by finding the distribution which when fitted to both distributions of mutations across sites, maximizes the likelihood. We then used this distribution to simulate data; we drew a random variate for each block from the distribution assigning this as the rate for that block. We then generated two Poisson variates with the appropriate means such that the total number of DNMs for each type of mutation was expected to be equal the total number of DNMs of those types.

5.4 Results

5.4.1 De novo mutations

To investigate large scale patterns of de novo mutations (DNMs) in humans we compiled data from four studies which between them had discovered 43,433 DNMs: 26,939 mutations from Wong *et al.* (2016), 11016 mutations from Francioli *et al.* (2016), 4931 mutations from Kong *et al.* (2012) and 547 mutations from Michaelson *et al.* (2012). We divided the mutations into 9 categorines reflecting the fact that CpG dinucleotides have higher mutation rates than non-CpG sites, and the fact that we

cannot differentiate which strand the mutation had occurred on: CpG C->T (a C to T or G to A mutation at a CpG site), CpG C->A, CpG C->G and for non-CpG sites C->T, T->C, C->A, T->G, C<->G and T<->A mutations. The proportion of mutations in each category in each of the datasets are given in table 5.1.

We find that the pattern of mutation differs significantly between the 4 studies (Chi-square test of independence on the number of mutations in each of the 9 categories, $p < 0.0001$). This appears to be largely due to the relative frequency of C->T transitions in both the CpG and non-CpG context. In the data from Wong *et al.* and Michaelson *et al.* the frequency of C->T transitions at CpG sites is ~13% whereas it is ~17% in the other two studies, a discrepancy which has been noted before between the studies of Michaelson *et al.* and Kong *et al.* (Eyre-Walker & Eyre-Walker 2014). For non-CpG sites the frequency of C->T transitions is ~24% in all studies except that of Wong *et al.* in which it is 26%. It is not clear whether these patterns reflect differences in the mutation rate between different cohorts of individuals, possibly because of age or race, or whether the differences are due to methodological problems associated with detecting DNMs. Since the differences are relatively small and it is not clear whether one study represents a more representative sample than the other, we combined the data.

Mutation type	Michaelson	Kong	Francioli	Wong	Overall
CpG C-T	0.128 (70)	0.173 (1816)	0.165 (3530)	0.131 (3530)	0.144 (6270)
CpG C-A	0.009 (5)	0.008 (37)	0.007 (79)	0.008 (204)	0.007 (325)
CpG C-G	0.009 (5)	0.007 (36)	0.007 (76)	0.006 (160)	0.006 (277)
Non C-T	0.236 (129)	0.237 (1169)	0.246 (2705)	0.264 (7121)	0.256 (11124)
Non T-C	0.28 (153)	0.267 (1319)	0.273 (3011)	0.283 (7636)	0.279 (12119)
Non C-A	0.086 (47)	0.083 (409)	0.084 (930)	0.089 (2399)	0.087 (3785)
Non T-G	0.073 (40)	0.073 (358)	0.073 (808)	0.071 (1920)	0.072 (3126)
Non C-G	0.097 (53)	0.085 (417)	0.081 (896)	0.088 (2372)	0.086 (3738)
Non T-A	0.082 (45)	0.067 (332)	0.063 (695)	0.059 (1579)	0.072 (2669)

Table 5.1. The proportion of mutations in each mutational category in each of the 4 datasets, number of mutations in parenthesis.

5.4.2 Distribution of rates

In thinking about large-scale variation in the mutation rate it is helpful to quantify the level of variation. This is not entirely straightforward since we have on average only 15 DNMs per Mb and hence a large level of sampling error. To overcome this problem we modelled the underlying variation in the mutation rate as a gamma distribution allowing mutations within each block to be Poisson distributed (i.e. the error variance was Poisson distributed), estimating the distribution using maximum likelihood (see chapter 2 for details of the method (Smith *et al.* 2016)). We fit a gamma distribution to the numbers of mutations dividing the genome into either 100kb or 1Mb windows. We do not consider scales below these because we have too little data - at 100kb we only have ~1.5 DNMs per region.

For the 1Mb data we find significant variation (i.e. the lower 95% confidence interval is not zero) when all mutations are considered together, however the level of variation is quite modest. We estimate the coefficient of variation (the standard deviation divided by the mean) of the distribution to be between 0.13 and 0.20 for most categories of mutation (Table 5.2). A gamma distribution with a coefficient of variation of 0.18 is a distribution in which only 10% of regions have a rate of mutation that is either 30% faster or 30% slower than the mean. The coefficient of variation is about twice as great at the 100kb level, but again this represents rather limited variation in the mutation rate; roughly 10% of regions have mutation rates that are either 60% slower or faster than the mean.

We also find significant variation for CpG transitions and non-CpG transitions and transversions (Table 5.2). However, we do not find significant variation for either CpG transversions, or when we split the data into individual mutational types; this is probably because we have too little data.

Mutation type	100kb	1Mb
All	0.40 (0.38, 0.41)	0.18 (0.17, 0.20)
CpG	0.35 (0.23, 0.45)	0.20 (0.13, 0.26)
nonCpG	0.24 (0.22, 0.27)	0.14 (0.12, 0.16)
CpG transitions	0.36 (0.22, 0.46)	0.19 (0.11, 0.26)
CpG transversions	0 (0, 0.95)	0.39 (0, 0.67)
nonCpG transitions	0.20 (0.15, 0.25)	0.13 (0.091, 0.15)
nonCpG transversions	0.29 (0.21, 0.35)	0.18 (0.14, 0.22)

Table 5.2. The coefficient of variation for a gamma distribution fitted to the density of DNMs, and the 95% confidence intervals of the coefficient of variation.

Given that there is variation in the rate of all mutational types, for which we have enough data, it is of interest to investigate whether the amount of variation differs between the mutational types. To investigate this we ran a series of likelihood ratio tests in which we fit separate and common distributions to the different mutational types. We found no significant differences in the amount of variation for CpG and non-CpG mutations or CpG transitions and CpG transversions (Table 5.3). However, we found evidence that there was significantly more variation for non-CpG transversions than transitions ($p<0.05$) (Table 5.3). Never-the-less, although significant, the differences in terms of the coefficient of variation are quite modest (Table 5.2).

Model	No of parameters	Log likelihood
100 kb		
Combined distribution for CpG and non-CpG	3	-54408.9
Separate distributions for CpG and non-CpG	4	-54407.3
Combined distribution for CpG transitions and transversions	3	-17634.5
Separate distributions for CpG transitions and transversions	4	-17634.44
Combined distribution for non-CpG transitions and transversions	3	-55347.3
Separate distributions for non-CpG transitions and transversions	**4**	**-55345.5**
1 Mb		
Combined distribution for CpG and non-CpG	3	-12022.2
Separate distributions for CpG and non-CpG	4	-12020.63
Combined distribution for CpG transitions and transversions	3	-6046.3
Separate distributions for CpG transitions and transversions	4	-6045.92
Combined distribution for non-CpG transitions and transversions	3	-12332.9
Separate distributions for non-CpG transitions and transversions	**4**	**-12330.03**

Table 5.3. Likelihood values for fitting combined and separate distributions to categories of mutations. Each pair of lines represents a likelihood ratio tests; bold figures denote a significant result.

5.4.3 Correlations between mutational types

Given that there is variation in the mutation rate at the 1Mb and 100kb levels and that this variation is quite similar for different mutational types, it would seem likely that the rate of mutation for the different mutational types are correlated. We find that this is indeed the case. At the 1Mb scale we find that the rate of CpG mutations is significantly correlated the rate of non-CpG mutations ($r = 0.086$, $p < 0.001$ by randomisation), and the rate of non-CpG transitions is correlated to the rate of non-CpG transversions ($r = 0.075$, $p < 0.001$). In both cases these correlations are about as strong as you would expect given the high level of sampling error; i.e. if we simulate data using a gamma distribution fit to both mutational categories, we find the mean correlation from 100 simulations are 0.080 and 0.083 respectively. In contrast we observe no significant correlation between the rates of CpG transitions and CpG transversions ($r = 0.018$, $p = 0.17$) but this appears to be due to the scarcity of CpG transversions, because the simulated data also show similarly low correlations (mean $r = 0.019$ from 100 simulations).

The pattern at the 100kb scale is slightly different. The rates of CpG and non-CpG mutations are significantly correlated to each other ($r = 0.013$, $p = 0.02$), as are the rates of non-CpG transitions and non-CpG transversions ($r = 0.036$, $p < 0.001$). However, whereas the correlation between non-CpG transitions and non-CpG tranversions is similar to the expected value (mean simulated $r = 0.030$) the correlation between CpG and non-CpGs is considerably smaller than the expected value (mean simulated $r = 0.031$) and less than 2% of simulated datasets have such a low correlation at the 100kb scale. The rate of CpG transversions is not correlated to the rate of non-CpG transversions ($r = 0.0084$, $p = 0.088$), which is consistent with simulated data (mean $r = 0.0069$), again reflecting the paucity of CpG transversions.

In summary there are significant correlations between the rates of CpG and non-CpG mutation, and non-CpG transitions and transversions, at both the Mb and 100kb scales, and these correlations are as strong as one would expect given the sampling error; this is with the exception of the correlation between CpG and non-CpG mutations at the 100kb scale, which is smaller than expected. There is no significant correlation between the rates of CpG transitions and CpG transversions, but this is as expected given the number of CpG transversions.

5.4.4 Correlations with genomic variables

To try and understand why there is large scale variation in the mutation rate, we compiled a number of genomic variables which have previously been shown to correlate to the rate of germline or somatic DNM, or divergence between species: recombination rates, GC content, replication time and nucleosome occupancy. All the variables are significantly correlated to the rate of DNM except replication time in the 1Mb analysis (Table 5.4). However, the lack of a correlation with replication time is deceptive, because if we run a multiple regression we find that all factors are significant including replication time. We also find that for almost all the individual mutational types the rate of DNM is negatively correlated to replication time (early replicating DNA has a lower mutation rate) (see appendix 5.1). A comparison of the standardised slopes suggests that nucleosome occupancy is the factor most strongly associated with the mutation rate followed by replication time and recombination rate.

Factor	100kb Correlation	100kb Slope	1Mb Correlation	1Mb Slope
GC content	0.053***	NS	0.097***	0.16
Recombination rate	0.071***	0.046	0.20***	0.11
Replication time	0.014*	-0.09	0.017	-0.22
Nucleosome occupancy	0.077***	0.12	0.15***	0.44

Table 5.4. The correlation coefficients from regressing DNM rate against individual features and the standardized slopes from including all significant features in a multiple regression. Note that a negative slope for replication time indicates that the mutation rate is higher for later replicating regions.

5.4.5 Correlation with divergence

The rate of divergence between species is expected to depend, at least in part, on the rate of mutation; patterns of substitution between species have commonly been used to infer patterns of mutation (Williams & Hurst 2000; Martin J. Lercher & Hurst 2002; Smith & Lercher 2002; M J Lercher & Hurst 2002; Hellmann *et al.* 2003; Tyekucheva *et al.* 2008; Duret & Arndt 2008; Stamatoyannopoulos *et al.* 2009; Don *et al.* 2013; Prendergast *et al.* 2007; Hodgkinson & Eyre-Walker 2011). However it is not known to what extent the variation in DNM density represents divergence. To investigate this we calculated the rate of divergence between humans and chimpanzees and correlated this to the denisity of DNMs in non-overlapping 1Mb windows. There are at least three different sets of human-chimpanzee alignments available, Ensembl EPO alignments (EPO), UCSC Pairwise (PW) and MultiZ 46-Way (MZ) alignments, and we find that the correlation between human DNM density and divergence depends upon the alignments used and the filtering applied to the windows (Figure 5.1). The correlation is significantly negative if we include all windows for the PW and MZ alignments, becoming more positive as we restrict the analysis to windows with more interrogable bases. In contrast the correlations are always positive when using the EPO alignments, and the strength of this correlation does not change once we get above 200,000 interrogable bases per 1Mb. Further analysis suggests there are some problems with the PW and MZ alignments; divergence per Mb window is inversely correlated to mean alignment length ($r = -0.31$, $p < 0.0001$) for the PW alignments and positively correlated ($r = 0.57$, $p < 0.0001$) for the MZ alignments (Appendix 5.2). The EPO alignments use a global alignment method (Paten *et al.* 2009), as opposed to local methods used for PW and MZ and as such can not exhibit such a bias. We consider these to be of the highest quality. Therefore we use the EPO alignments for the rest of this analysis. Figure 5.2 shows the correlation between DNM frequency and divergence for the EPO alignments.

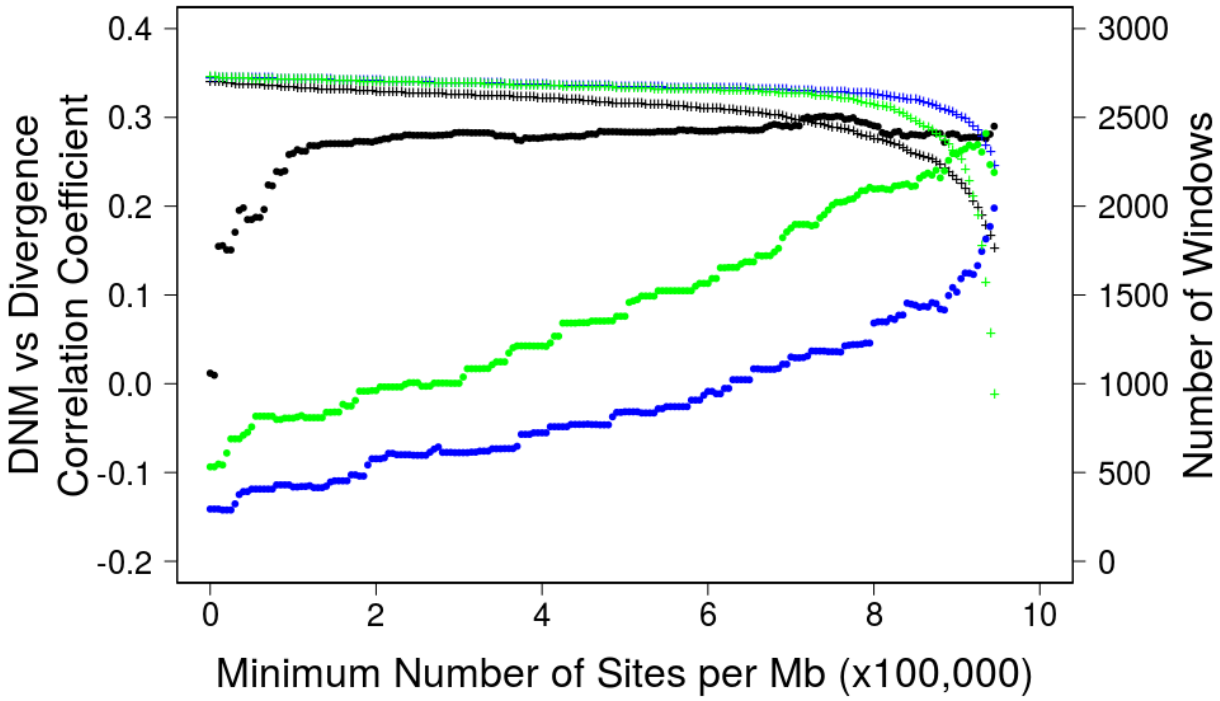


Figure 5.1. Correlation Coeffiecients of DNM density versus divergence as a function of the minimum number of interrogable sites per Mb window in steps of 5000 bases. The result for 3 alignment methods are shown; PW (blue circles), MZ (green circles) and EPO (black circles). Crosses denote the number of windows remaining when each filter is applied.

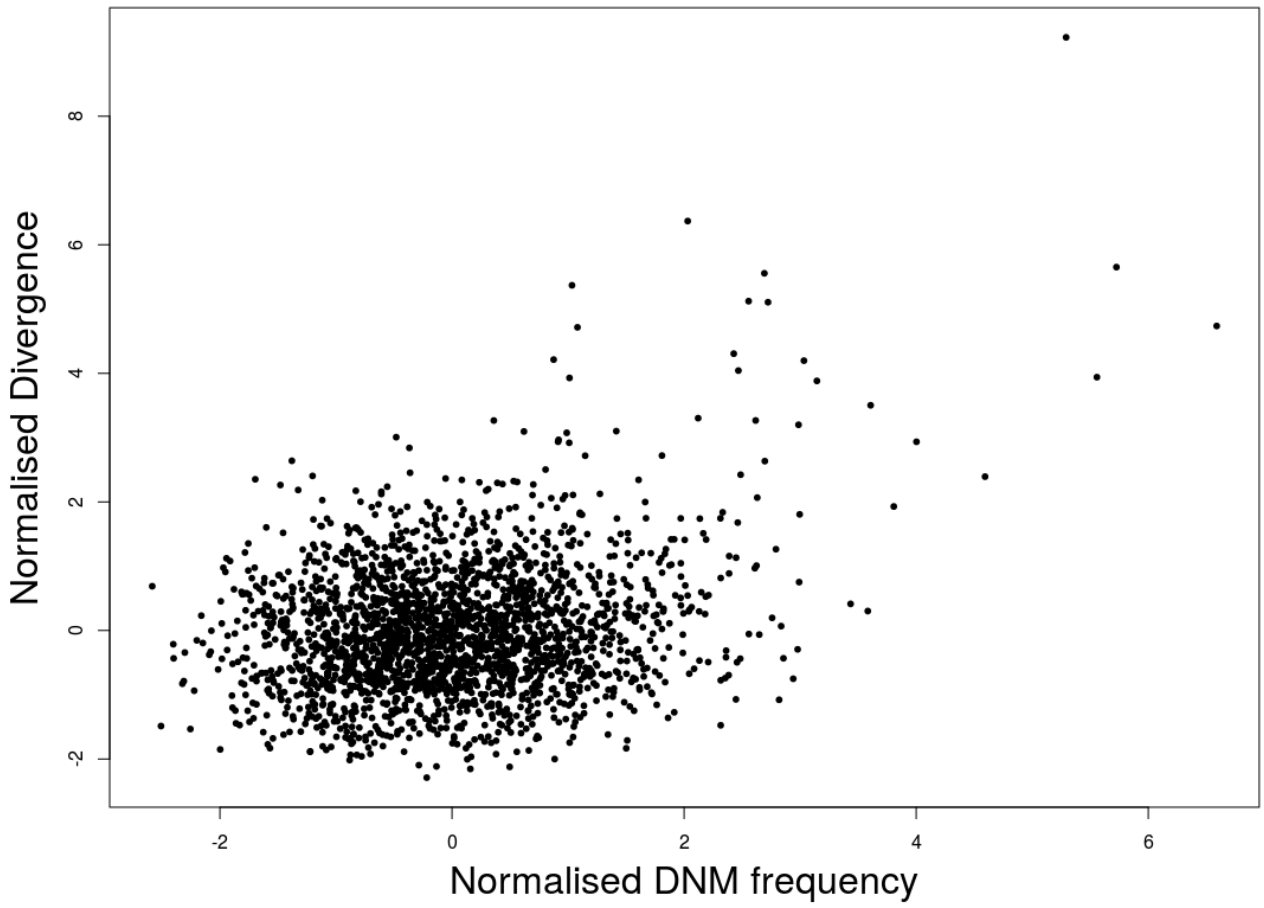


Figure 5.2. Graph showing the correlation of DNM frequency and divergence for the EPO alignments at 1Mb, filtered for windows only containing >500,000 interrogable sites (2,337 1Mb windows) with a correlation coefficient of 0.24. DNM frequency and divergence has been scaled, the axes represent units of standard deviation.

The EPO alignments allow us to consider lineage specific changes using parsimony to reconstruct ancestral states. As expected the divergence along the human lineage is correlated to the rate of DNM (0.24 at the 1Mb scale, 0.064 at the 100kb scale, $p < 0.0001$). However, the correlation between the rate of DNMs and divergence is not expected to be perfect even if variation in the mutation rate is the only factor affecting the rate of substitution between species; this is because we have relatively few DNMs and hence our estimate of the density of DNMs is subject to a large amount of sampling error. To investigate how strong the correlation could be, we follow the procedure suggested by Francioli *et al.* (2015); we assume that variation in the mutation rate is the only factor affecting the substitution rate between species and that we know the substitution rate without error (this is an approximation, but the sampling error associated with the substitution rate is small relative to the sampling error associated with DNM density). We then simulate the observed DNMs for each window using the rate of substitution for each window as the mean rate for the poisson process, and then consider the correlation between these simulated DNMs and the observed substitution rates. We repeated this procedure 10,000 times to generate a distribution of expected correlations. Doing this we find that the correlation between divergence and DNM density could be 0.50 at the 1Mb level and 0.24 at 100kb level, considerably greater than the observed values of 0.24 and 0.064 respectively (Table 5.5). In none of the simulations was the simulated correlation lower than 0.44 for 1Mb and or 0.21 for 100kb. Similar patterns hold for almost all mutational types (Table 5.5).

Mutation	100kb Obs. correlation	100kb Exp. correlation	1Mb Obs. correlation	1Mb Exp. correlation
All	0.064	0.24	0.24	0.5
CpG C-T	0.055	0.11	0.2	0.21
CpG C-A	0.021	NA	0.053	0.086
CpG C-G	0.011	NA	0.033	0.085
non C-T	0.016	0.14	0.063	0.29
non C-A	0.047	0.12	0.16	0.27
non T-C	0.015	0.14	0.08	0.32
non T-G	0.0099	0.096	0.04	0.21
non C-G	0.059	0.11	0.23	0.26
non T-A	0.013	0.091	0.094	0.18

Table 5.5. The observed and expected correlations between the density of DNMs and substitutions at the 100kb and 1Mb scales; the expected correlation is the mean correlation from 10,000 simulations assuming that all the variation in the substitution rate is due to variation in the mutation rate (and assuming the pattern of mutation has not changed along the human lineage). We are not able to simulate data for CpG transversions due to the fact that some regions have no substitutions of this type.

5.4.6 The effect of recombination

There are several potential explanations for why the correlation is weaker than it could be; the pattern of mutation might have changed, or there might be other factors that affect divergence. Francioli *et al.* (2015) showed that including recombination in the regression model significantly improved the correlation between divergence and DNM density; a result we confirm here; the square root of coefficient of determination when recombination is included in a regression of divergence versus DNM density increases from 0.24 to 0.43, and from 0.064 to 0.22 for the 1Mb and 100kb datasets respectively. As detailed in the introduction there are at least four explanations for why including recombination in the regression might improve the correlation: (i) BGC can affect the probability of a mutation reaching fixation, countering S>W mutations and driving the fixation of W>S mutations, thus substitutions in regions of high recombination may reflect the processes of both BGC and de novo mutation. (ii) Recombination can also affect the probability that a mutation will be fixed by natural selection; in regions of high recombination deleterious mutations are less likely to be fixed, whereas advantageous mutations are more likely, and so the pattern of substitution may reflect both selection as influenced by recombination and de novo mutation. (iii) Recombination can affect the depth of the genealogy in the human-chimpanzee ancestor; regions of low recombination can increase the effects of hitch-hiking and background selection, resulting in a lower effective population size and thus less diversity in the ancestor. As ancestral polymorphisms segregate into the human and chimpanzee branches, regions of low recombination will therefore have a fewer ancestral polymorphisms contributing to divergence between the two species than regions of higher recombination. (iv) Problems with regressing against correlated variables that are subject to sampling error. For example, let us imagine that factor X (de novo mutation) affects diversity directly whereas factor Y (recombination rate) does not; however, X and Y are mildly correlated - This could be the case if recombination is mutagenic but does not exhibit BGC. If we

can measure X and Y without error then a multiple regression should show that diversity is correlated to X but not Y. However, if we cannot measure X without error then we may find that the rate of diversity correlates to both X and Y.

We can potentially differentiate between these four explanations by comparing the slope of the regression between the rate of substitution and recombination rate, and the rate DNM and recombination rate. If recombination affects the substitution rate independent of its effects on DNM mutations because of BGC, then we expect the slope associated with the substitution rate to be greater for AT->GC, smaller for GC->AT, and unaffected for G<->C and A<->T changes. If selection is the cause then we expect all the slopes associated with substitutions to be less than those associated with DNMs. The reason is as follows; if a proportion of mutations are slightly deleterious then those will have a greater chance of being fixed in regions of low recombination than high recombination. If the effect of recombination is due to variation in the coalescence time in the human-chimp ancestor, then we expect all the slopes associated with substitution to be greater than those associated with DNMs; this because the average time to coalescence is expected to be shorter in regions of low recombination than in regions of high recombination. Finally, if the effect is due to problems with multiple regression then we might expect all the slopes to become shallower. Since the DNM density and divergences are on different scales we divided each by their means to normalise them and hence make the slopes comparable.

As Francioli *et al.* (2015) noted, the overall rate of DNM is positively correlated to the rate of recombination, suggesting that recombination might be mutagenic (Tables 5.4, 5.6). The average mutation rate in the decile with the highest recombination rate is approximately 20% higher than the mutation rate in the lowest decile (Figure 5.3). We are able to show in addition that the rate of DNM is correlated to the recombination rate at both the 1Mb and 100kb scales and for almost all

mutational types; the exceptions are non-CpG T->G and T<->A at both scales, and non-CpG C->A at the 1Mb scale (Table 5.6).

	100kb			1Mb		
	DNM v RR	**Div v RR**	**Difference test p-value**	**DNM v RR**	**Div v RR**	**Difference test p-value**
Total	0.042***	0.028***	0.13	0.069***	0.066***	0.38
CpG C>T	0.039**	-0.005***	0.09	0.043*	-0.009	0.001
C>T	0.037***	0.006***	0.09	0.063***	0.016***	<0.001
C>A	0.042**	-0.004**	0.11	0.036	-0.014*	0.001
T>A	0.022	0.010***	0.4	0.041	0.026***	0.28
T>C	0.019**	0.042***	0.15	0.035**	0.096***	<0.001
T>G	0.007	0.044***	0.18	0.013	0.106***	<0.001
C>G	0.061***	0.027***	0.22	0.090***	0.076*	0.26

Table 5.6. Slope of the regression of the divergence along the human lineage, since the split with chimpanzees, and recombination rate, and the slope of the regression of the DNM rate and recombination rate for 100kb and 1Mb windows, and the result of testing whether the slopes are significantly different, using bootstrapping. * p<0.05, ** p<0.01, *** p<0.001.

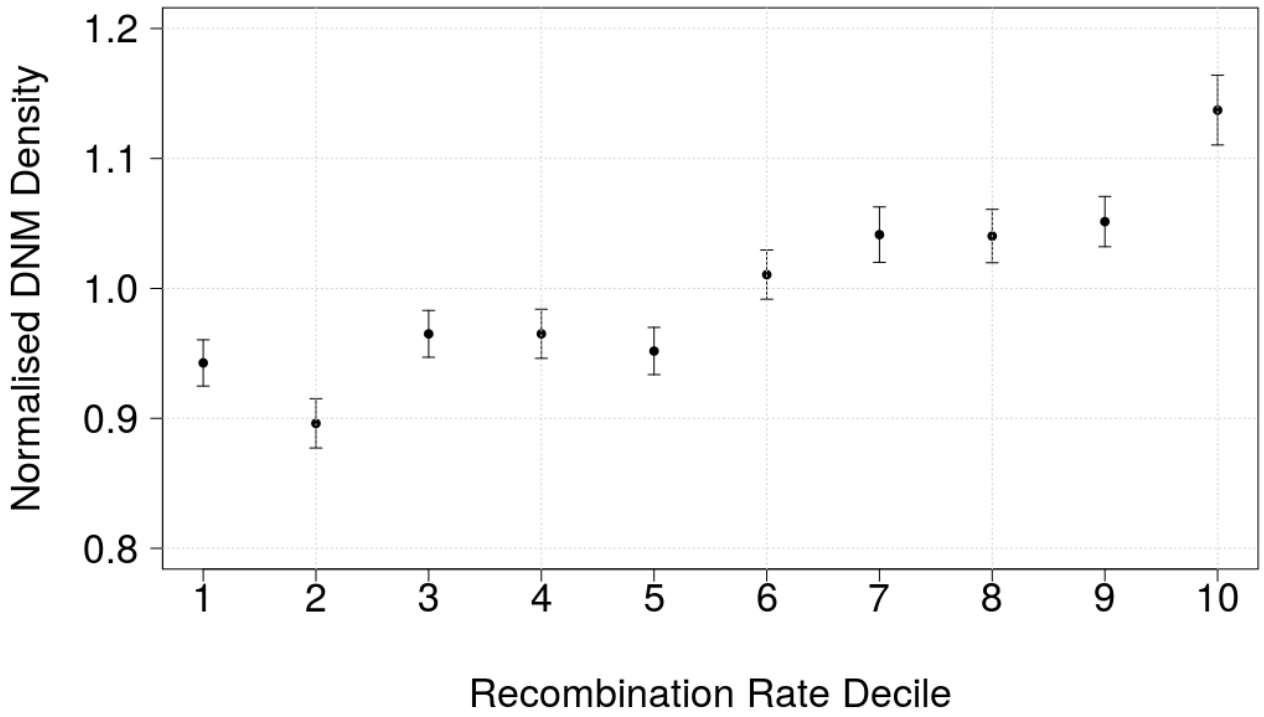


Figure 5.3. Normalised DNM density for each recombination rate decile from lowest (1) to highest (10). Vertical lines represent the standard error.

In terms of our test of why the substitution rate depends upon the recombination rate independent of its effects on the mutation rate, we find that the slopes are less positive, and sometimes even negative, for all GC->AT changes and more positive for AT->GC changes (we do not consider CpG transversions because there are too few DNMs) (Table 5.6). These differences in the slopes, as determined by bootstrapping, are significant at the 1Mb scale but non-significant at the 100kb scale. We find no significant difference in the slopes associated with A<->T and G<->C at either scale. These results are highly consistent with BGC affecting the rate of substitution between species.

5.4.7 Other species

Divergence between species, usually humans and macaques, is often used to control for mutation rate variation in various analyses (Gossmann *et al.* 2011; Burgess & Yang 2008; McVicker *et al.* 2009). But how does the correlation between divergence and the DNM rate in humans change as the species being compared get further apart? To investigate this we compiled data from a variety of primate species – human/chimpanzee/orang-utan (HCO) considering the divergence along the human and chimp lineages, human/orang-utan/macaque (HOM) considering the divergence along the human and orang-utan lineages, and human/macaque/marmoset (HMM) considering the divergence along the human and macaque lineages. This yields two series of divergences of increasing evolutionary divergence: the human lineage from HCO, HOM and HMM, and chimp from HCO, orang-utan from HOM and macaque from HMM. All divergences were normalised by dividing by their mean. For both series we see a clear tendency for the slope of the regression between divergence and DNM rate to decrease as a function of evolutionary divergence at both the 100kb and 1Mb scales (Figure 5.4). If we calculate the correlation coefficient between the slope and the evolutionary stratum, assigning 1, 2 and 3 to the strata (e.g. 1 for chimp, 2 for orangutan and 3 for macaque), we find that the correlations are negative for all mutational types, for both sets of

evolutionary divergence and scales (binomial test of positive versus negative for both 100kb and 1Mb $p < 0.01$) (Figure 5.4).

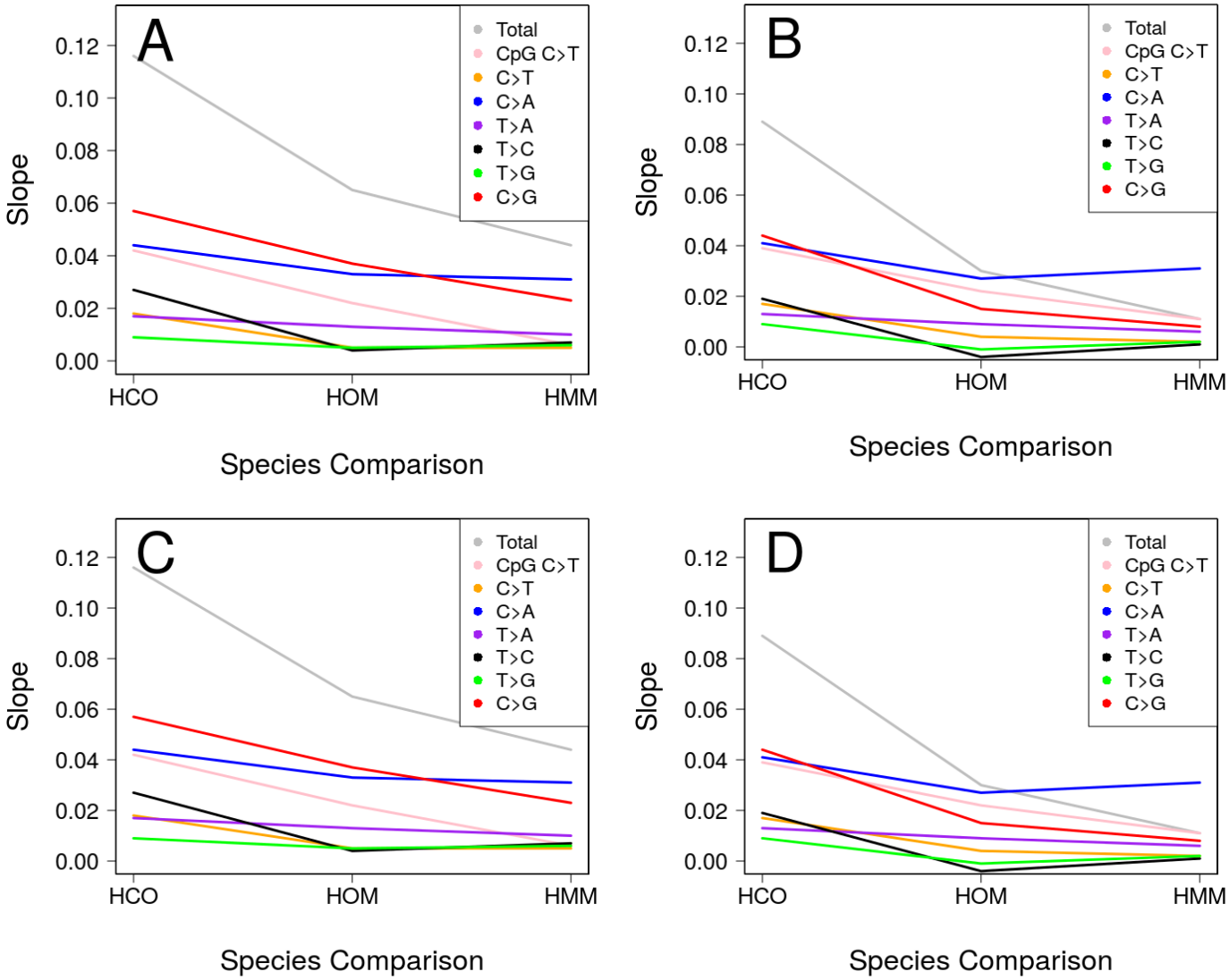


Figure 5.4. The slope of the linear regression between divergence and DNM rate for 1Mb (top panels; A and B) and 100kb (bottom panels; C and D). HCO is the divergence since humans split from chimpanzee, from a comparison of human, chimpanzees and orang-utans; HOM is the human divergence since humans split from orang-utans, using human, orang-utan and macaque; HMM is the human divergence since humans split from macaques using human, macaque and marmoset). The left panels; A and C are for substitutions specific to the human lineage, the right panels; B and D are for substitutions specific to the chimpanzee, orang-utan and macaque lineages.

5.4.8 Correlation with diversity

Just as we expect there to be correlation between divergence and DNM rate, so we might expect there to be correlation between DNA sequence diversity within the human species and the rate of DNM. To investigate this we compiled the number of SNPs in 1Mb and 100kb blocks from the 1000 genome project (The 1000 Genomes Project Consortium 2015). There is a positive correlation between SNP density and DNM rate at both the 1Mb ($r = 0.36$, $p<0.001$) and 100kb scales ($r = 0.13$, $p<0.001$).

Using a similar strategy to that used in the analysis of divergence we calculated the correlation we would expect if all the variation in diversity was due to variation in the mutation rate by assuming that the level of diversity was known without error, and hence was a perfect measure of the mutation rate (we have on average 31,000 SNPs per 1Mb, so there is little sampling error associated with the SNPs). We then simulated the observed number of DNMs according to these inferred mutation rates. The expected correlations are 0.41 at the 1Mb and 0.17 at the 100kb scales, significantly ($p<0.01$ in both cases), but only slightly greater than the observed correlations. This is consistent with the role of BGC affecting the fate of new mutations, because just as with selection, BGC is expected to affect the probability of fixation more than the probability of contributing to diversity. To illustrate this we can use the following reasoning; for S>W mutations in areas of high recombination, BGC will work against the derived W allele to fix the ancestral S allele. The derived W allele may persist for some time in the population but will ultimately be removed by BGC, thus BGC has a greater effect on divergence than diversity, effectively reducing divergence for S>W substitutions in areas of high recombination. For W>S mutations, in areas of high recombination BGC favours the derived allele, S, effectively pushing it quickly through the population to fixation, again having a greater effect on divergence over diversity, here increasing divergence for W>S

substitutions. In both cases, whether the result is an increase or decrease divergence for highly recombining areas, we can see that divergence is more sensitive to the force of BGC than diversity.

5.5 Discussion

We have considered the large-scale distribution of DNMs along the human genome and the relationship between the rate of DNM, divergence between species, and diversity within a species. We find that there is significant variation in the mutation rate at both the 1Mb and 100kb scales (we cannot investigate smaller scales due to a lack of DNMs). However, the variation in the mutation rate is quite modest; at the 1Mb scale 90% of regions have a mutation rate that is within ± 30% of the mean, at the 100kb scale this increases to ± 60% of the mean. It seems likely that there will be more variation at smaller scales.

Although we do not have enough DNM data to consider each mutational type individually it is evident that the rate of CpG and non-CpG mutations varies across the genome as does the rate non-CpG transitions and transversions at both the 1Mb and 100kb scales; we do not have enough data to determine what is happening with CpG transversions. The rate of mutation of the different mutational types are about as strongly correlated to each other as they could be suggesting that they are influenced by similar factors. That is, the correlation we observe between different mutational types, say CpG and non-CpG mutations, is very close to the correlation we get when simulating DNMs for CpG and non-CpG mutations from the same distribution - see section 5.3.6 for methods.

We confirm that replication time, recombination rate and GC content are all independently correlated to the rate of DNM, however the strongest effect we find comes from nucleosome

occupancy. Although, nucleosome occupancy has previously been investigated, no significant effects were detected at these scales (Michaelson *et al.* 2012). However, this study only considered a small number of DNMs and at smaller scales of 100bp it did infer a significant effect of nucleosome occupancy. As nucleosome occupancy is intimately related to the secondary DNA structures, this may support a role for chromatin architecture in the determination of mutation rates at large scales, as has previously been postulated (Makova & Hardison 2015; Don *et al.* 2013).

As expected the rate of divergence between species is correlated to the rate of DNM, however, as Francioli *et al.* (2015) showed the correlation is worse than it would be if variation in the mutation rate was the only factor affecting divergence. They showed that although the rate of DNM is correlated to the rate of recombination, as might be expected if recombination is mutagenic, that divergence is also correlated to the rate of recombination independently of this effect, as you might expect from the effect of BGC. We have shown that the reason recombination affects divergence, independent of its effect on mutation, is indeed likely to be due to effect of BGC, since the slope of the relationship between divergence and recombination rate is smaller than the slope for DNM rate and recombination for GC->AT changes, but greater for AT->GC changes; A<->T and G<->C changes are unaffected.

We also show that the relationship between divergence and DNM rate gets weaker (the slopes get shallower) as more and more divergent species are considered. This might be due to two factors. First we might expect the mutation rate of a region to evolve through time eroding the relationship between divergence and the current mutation rate. Second, the relationship might get weaker because we are underestimating the divergence as species get more divergent. This might tend to affect more divergent blocks the most. However, we see no obvious evidence of this; the mutation that should be most affected is CpG transitions and the decay in the slope (between divergence and

DNM rate) is no faster than for other mutational types (Figure 5.4).

Divergence between species has often been used to control for mutation rate variation in humans - see (Gossmann *et al.* 2011; Burgess & Yang 2008; McVicker *et al.* 2009). This is clearly not satisfactory given that divergence is more strongly correlated to the rate of recombination than the rate of DNM, and the relationship between divergence and the rate of DNMs decreases as evolutionary divergence increases. In the coming years it is likely that many more thousands of DNMs will be discovered and a high-resolution map of the rate of mutation across the human genome will be produced. In the meantime however it seems that diversity within humans is a better proxy for mutation rate variation than divergence, since the correlation between diversity and divergence is about as good as it can be. Never-the-less two questions remain unresolved. Over what scale does the mutation rate in humans vary? And if its at a smaller scale than 100kb, does diversity still mirror this variation accurately?

It has been known for sometime that diversity across the human genome is correlated to the rate of recombination (Martin J. Lercher & Hurst 2002; Hellmann *et al.* 2003; Hellmann *et al.* 2005) and there has been much debate about whether this is due to mutagenic effects of recombination or the effect of recombination on processes such as genetic hitch-hiking and background selection. Divergence between humans and other primates is correlated to the rate of recombination, which was initially interpreted as being due to a mutagenic effect of recombination, but subsequently it has been interpreted as evidence of BGC. Both of these hypotheses appear to be correct – as we have shown here and as Francioli *et al.* (2015) show, the rate of DNM is correlated to the rate of recombination, but the recombination rate also affects which mutations become fixed through BGC. However, it has also been shown that there is a correlation between diversity and recombination even when divergence is controlled for (Hellmann *et al.* 2005) and this has led to the suggestion that

diversity is also affected by linked selection in humans. If we run a multiple regression of SNP density against the rate of DNM and recombination we find that both factors are significant at both the 1Mb and 100kb scales. However, the independent effect of recombination could simply be due to the fact that both the DNM rate and recombination rate are measured with error. An illustration of this probleem is given in section 5.4.6. This demonstrates that recombination could appear to be independently correlated to SNP density, even though its only effect is on the mutation rate. Our analysis of the correlation between SNP density and the rate of DNMs would suggest that there is little or no independent effect of recombination impacting diversity within humans because the correlation is almost as strong as it could be.

6. Discussion and Conclusions.

6.1 The importance of mutations

The rate of mutation is important in the contexts of both evolution and disease. In evolutionary studies, it is central to the concept of a molecular clock and important for phylogenetic dating, from which much of our understanding of our evolutionary history is derived. Deleterious mutations are the cause of many diseases, including cancers, so it follows that the rate of mutation at a disease causing locus will affect the prevalence of the disease. Thus variation in the mutation rate along the human genome, if not properly understood and accounted for, has the potential to confound evolutionary studies, produce spurious driver candidates in cancer studies, and hinder the diagnostics aimed at understanding the etiologies of genetic diseases.

6.2 The key contributions of this thesis

In this thesis I considered mutation rate variation in both the germ-line and somatic tissues, at varying scales. In the germ-line, the major advance of this thesis over previous works is use of direct methods to analyse the magnitude, scale and determinants of mutation rate variation. This has enabled us to tease apart the evolutionary and mutational forces, whilst directly quantifying the variation in the human mutation rate at different scales; at large scales the variation appears to be quite modest, however at the single nucleotide scale there is potentially huge cryptic variation in the mutation rate. I have also sought to extend this work into somatic tissues, however due to the quality of data and heterogeneity of samples and cell types, the primary findings lean towards

highlighting areas of improvement for NGS pipelines, and the development of methods with which future studies could provide more insight. With these methods and the ever increasing flow of somatic SNVs, coupled with the continual improvements in NGS technology, it should soon be possible to provide accurate answers to the questions posed of somatic mutation rate variation in this thesis.

6.3 Direct measurements of human mutation rate variation

In chapter 2, I use DNMs from disease genes to quantify the variation in the mutation rate at the single nucleotide level. In doing so I present the first direct estimate of cryptic variation in the germ-line, a phenomenon only previously inferred from comparative methods (Johnson & Hellmann 2011; Hodgkinson *et al.* 2009). For all mutations excluding the hyper-mutable CpG transitions, the best estimate indicates that cryptic variation in the mutation rate varies hundreds to thousands of fold between sites. This far exceeds the estimates from previous comparative studies by at least one order of magnitude (Johnson & Hellmann 2011; Hodgkinson & Eyre-Walker 2011). By contrast, the hyper-mutable CpG transitions exhibit very little variation in their mutation rate. However their overall rate of mutation relative to non-CpG transversions appear to be much higher in these disease genes (90 - 640 fold), when compared to the whole genome (19 to 28 fold). We must be cautious when interpreting these results. Although mutation rate variation is the most likely origin of the heterogeneity, and every effort has been made to remove other causes, such as variable penetrance, ascertainment bias or alternative splicing, we can not entirely exclude these possibilities. In addition, the sample size is relatively small, and so only allows the analysis of four mutation types; CpG, non-CpG, transition and transversion, limiting the power of the analysis.

In chapter 5, I use ~43,000 DNMs from whole genome sequenced pedigrees, which enabled the the direct quantification of mutation rate variation at larger scales. I showed that the variations in mutation rate at the 1Mb and 100kb scales are modest, with the ratio of the rates of the first and last deciles being 2 and 3 respectively. These rates are significantly below those found at the single nucleotide level, in terms of both cryptic variation described in chapter 2 and in context dependent variation from other direct studies (Kong *et al.* 2012; Michaelson *et al.* 2012). It demonstrates the need to consider different determinants of variation at different scales; it is clear that most of the variation that does exist is at the fine scale, and although there is variation at large scales, it can not all be explained in terms of fine scale variation. It is hoped that in the near future enough DNMs will be available to extend this analysis to the 10Kb scale, where I would predict we would find greater than the 3-fold variation that we see at the 100kb scale.

6.4 Comparative versus direct methods

Much of our inferences about mutation rates has come from investigating patterns of divergence; the assumption usually being that only a small fraction of the genome is under selection, thus patterns in the "neutral" parts of the genome should broadly represent the patterns of mutation and maybe other non-selective evolutionary processes. The recent release of ~43,000 DNMs has allowed, for the first time with sufficient power, a direct comparison of the patterns of DNM density, divergence and diversity. Surprisingly, in chapter 5, I show that the pattern of divergence does not actually represent the pattern of mutation very well. The variation in DNM density has a correlation coefficient of only half of what we would expect with the power we have, in agreement with previous work (Francioli *et al.* 2015). My analysis suggests that this is due to the process of BGC. In contrast, variation in the rate of mutation seems to explain almost all the variation there is

in diversity, at least at the scales I have considered. Interestingly diversity and divergence are highly similar in their patterns of substitution; variation in diversity can explain 62% of the variance in divergence, suggesting the processes shaping our genomes beyond that of mutation work on small evolutionary timescales, within a species. Furthermore, when regressing substitution density between humans and other primates of increasing evolutionary distance against DNM density, we note that although the relationship weakens, it remains relatively strong even with species as diverged as marmoset. This suggests that although these evolutionary processes may act upon our genomes over short timescales, their footprints remain apparent for a long time.

Much attention has been given to genomic features that are thought to influence mutation rates, such as GC content (Hellmann *et al.* 2005; Wolfe *et al.* 1989; Tyekucheva *et al.* 2008; Don *et al.* 2013; Chen *et al.* 2010) and replication time (Chen *et al.* 2010; Pink & Hurst 2010; Stamatoyannopoulos *et al.* 2009; Don *et al.* 2013) and those involved in chromatin architecture (Polak *et al.* 2015; Makova & Hardison 2015). Replication time is one of the strongest correlates in both comparative genomics and somatic studies and the evidence presented in chapter 5, suggests that it is fairly strongly associated with DNM density for most mutational types. However, the association appears to be stronger with divergence than DNM density. Out of the features I investigated, nucleosome occupancy, which is linked to chromatin organisation (Makova & Hardison 2015; Don *et al.* 2013), appears to be have the strongest association to DNM density. This has previously only been shown to be to be correlated at much smaller scales (Michaelson *et al.* 2012), and so chapter 5 provides the first direct evidence that determinants of chromatin architecture at large scales influence mutation rate.

6.5 Somatic mutation rate variation

When more than 3 million SNVs from over 500 cancers containing many recurrently hit sites were publicly released (Alexandrov *et al.* 2013), it presented a great opportunity to apply the methods from chapter 2 to somatic SNVs, to produce the first quantification of cryptic variation in somatic tissues. The ~1000 mutations per Mb allowed enough power to comprehensively control for the effect of neighbouring nucleotides on the mutation rate. However, when many of the recurrently hit sites appeared to be artifactual, it prompted an extension to the method, to try and quantify the level of error in the data. I concluded that 2-4% of SNVs were the product of error. With the rate of error this high, it was not possible to accurately quantify the level of cryptic variation. However, this finding was important in highlighting the highly site specific, yet cryptic error emanating from NGS pipelines. It demonstrates that recurrence of mutation, often used to detect drivers of cancers or disease causing variants, can not be taken alone as evidence of causality. Other considerations, such as the mappability of a region, or validation with other methods should also be sought. The accuracy of NGS technology is also continually improving, so as cohorts of whole genome sequenced cancers increase in size - such as the recently released 560 breast cancers (Morganella *et al.* 2016) - improving the homogeneity of sample processing and NGS work-flows, it is foreseeable that the aim of accurately quantifying the cryptic variation in somatic tissues will soon be possible.

Although chapter 3 estimated 2-4% of SNVs were errors, these were highly site specific, and thus only significantly affect the results of analyses at the single nucleotide scale. Therefore, in chapter 4, using the same data chapter 3, the aim was to investigate the relationship of SNVs with fragile sites (FS), a phenomenon associated with larger scales. This had not been attempted before, despite both SNVs and FSs being linked to genomic instability and potentially sharing a common biology in their mechanism of formation. To our surprise we could find no strong evidence linking the

distribution of SNVs to FSs. However, there were many challenges in designing this study which may have obscured possible associations. Owing to the various resolutions to which FSs have been previously mapped and the various criteria used to determine a site as fragile, classifying a genomic region as a FS or a non-FS was not straight forward. Adding to this, the heterogeneity of cancer samples used, cell lines available for replication time data and cell types in which FS were detected, it is possible that there are subtle associations still to be found. As with the outcome of chapter 3, chapter 4 has laid the ground for further studies in the not too distant future, where increased cohort sizes will provide better sample homogeneity which should allow for both greater sensitivity and specificity.

6.6 future directions

This thesis has demonstrated the importance of being able to analyse the DNMs directly as opposed to inferring patterns of DNM from comparative methods. There are currently ~43,000 DNMs available from pedigrees, so it is not possible to directly study the magnitude of variation in the germ-line below the 100kb scale. However this number is growing quickly and predicted to exceed 100,000 by the end of 2016 (Wendy Wong, private communication). Thus it will soon be possible to take the next logical step, and expand the work of chapter 5 to examine smaller windows, where I would expect the magnitude of mutation rate variation would increase. Increasing the resolution of this analysis in this way, combined with a more exhaustive interrogation of associated genomic features, should provide a deep insight into both the true determinants of mutation rate variation, and the processes driving patterns of divergence. One the most interesting insights from this thesis is the apparent disconnect between DNM density and divergence. This may suggest that many of the proposed determinants of mutation rate, such as replication time and chromatin architecture,

may have a greater influence on evolutionary processes than mutation. This is a particularly interesting hypothesis if we are to also consider patterns of somatic mutation. SNVs from cancer genomes are generally treated as somatic DNMs, in that they arose in the somatic tissue of the individual over a relatively short timescale - years as opposed to the millions of years of evolutionary time. Selection is also thought to be relaxed in somatic tissues (Yadav *et al.* 2016), so very little credence has been given to the possibility that patterns of somatic mutation may actually be dominated by evolutionary processes. However, if we consider, (i) evolutionary time in terms of the number cell divisions as opposed to absolute time, (ii) that clonal selection is highly prevalent in cancers and, (iii) that cell proliferation in many cancers is rampant, then it is plausible that the patterns of somatic mutation may strongly reflect evolutionary pressures and therefore be more representative of divergence as opposed to germ-line DNM. This may also explain why a feature such as replication time exhibits considerable variation with somatic SNVs and divergence, but has a lesser impact on the rate of DNMs. The continual proliferation of SNV/DNM data coming from trio studies, cancer sequencing studies and initiatives such as the 100k genomes project over the next few years, should enable a shift away from the reliance on comparative methods to detect determinants of mutations, and result in a much deeper understanding of the forces of mutation and evolution in both the germ-line and the soma.

References

Aguilar, A., Smith, T.B. and Wayne, R.K., 2005. A comparison of variation between a MHC pseudogene and microsatellite loci of the little greenbul (Andropadus virens). BMC Evolutionary Biology, 5(1), p.47.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L. and Boyault, S., 2013. Signatures of mutational processes in human cancer. Nature, 500(7463), pp.415-421.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S. and Stratton, M.R., 2015. Clock-like mutational processes in human somatic cells. Nature genetics, 47(12), pp.1402-1407.

Ananda, G., Chiaromonte, F. and Makova, K.D., 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. Genome biology, 12(3), p.R27.

Araten, D.J., Golde, D.W., Zhang, R.H., Thaler, H.T., Gargiulo, L., Notaro, R. and Luzzatto, L., 2005. A quantitative measurement of the human somatic mutation rate. Cancer research, 65(18), pp.8111-8117.

Arlt, M.F., Ozdemir, A.C., Birkeland, S.R., Lyons, R.H., Glover, T.W. and Wilson, T.E., 2011. Comparison of constitutional and replication stress-induced genome structural variation by SNP array and mate-pair sequencing. Genetics, 187(3), pp.675-683.

Averof, M., Rokas, A., Wolfe, K.H. and Sharp, P.M., 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science, 287(5456), pp.1283-1286.

Awadalla, P., Gauthier, J., Myers, R.A., Casals, F., Hamdan, F.F., Griffing, A.R., Côté, M., Henrion, E., Spiegelman, D., Tarabeux, J. and Piton, A., 2010. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. The American Journal of Human Genetics, 87(3), pp.316-324.

Barron, V.A. and Lou, H., 2012. Alternative splicing of the neurofibromatosis type I pre-mRNA. Bioscience reports, 32(2), pp.131-138.

Baskaran, S. & Brahmachari, V., 2000. Chromosomal Fragility and Human Genetic Disorders. Indian Journal of Clinical Biochemistry, 15(Supplement 1), pp.145–157.

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G. and Tarpey, P.S., 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature, 513(7518), pp.422-425.

Beletskii, A & Bhagwat, A.S., 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America, 93(24), pp.13919–13924.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research 27:573–580.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M. and Mc Henry, K.T., 2010. The landscape of somatic copy-number alteration across human cancers. Nature, 463(7283), pp.899-905.

Bhattacharyya, N.P., Ganesh, A., Phear, G., Richards, B., Skandalis, A. and Meuth, M., 1995. Molecular analysis of mutations in mutator colorectal carcinoma cell lines. Human molecular genetics, 4(11), pp.2057-2064.

Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C. and Widaa, S., 2010. Signatures of mutation and selection in the cancer genome. Nature, 463(7283), pp.893-898.

Bird, A.P., 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research, 8(7), pp.1499–1504.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D. and Haussler, D., 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome research, 14(4), pp.708-715.

Bulmer, M., 1986. Neighboring base effects on substitution rates in pseudogenes. Molecular biology and evolution, 3(4), pp.322–329.

Burgess, R. & Yang, Z., 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Molecular Biology and Evolution, 25(9), pp.1979–1994.

Chen, C.L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O. and Thermes, C., 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome research, 20(4), pp.447-457.

Chen, J.M., Cooper, D.N. & Ferec, C., 2014. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: ~0.4% of the single-nucleotide substitution mutation rate. Human Mutation, 35(3), pp.392–394.

Chiaromonte, F., Yap, V.B. & Miller, W., 2002. Scoring pairwise genomic sequence alignments. Pacific Symposium on Biocomputing, 126, pp.115–126.

Christodoulou, J., Grimm, A., Maher, T. and Bennetts, B., 2003. RettBASE: the IRSA MECP2 variation database—a new mutation database in evolution. Human mutation, 21(5), pp.466-472.

Cohen, N.M., Kenigsberg, E. & Tanay, A., 2011. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell, 145(5), pp.773–786.

Conrad, D.F. *et al.*, 2011. Variation in genome-wide mutation rates within and between human families. Nature genetics, 43(7), pp.712–4.

Cooper, D.N. & Krawczak, M., 1990. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Human Genetics, 85(1), pp.55–74.

Coulondre, C. *et al.*, 1978. Molecular basis of base substitution hotspots in Escherichia coli. Nature, 274(5673), pp.775–780.

Debatisse, M., Le Tallec, B., Letessier, A., Dutrillaux, B. and Brison, O., 2012. Common fragile sites: mechanisms of instability revisited. Trends in Genetics, 28(1), pp.22-32.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A. and Malkova, A., 2011. Break-induced replication is highly inaccurate. PLoS Biol, 9(2), p.e1000594.

Dekaban, A., 1965. Persisting Clone of Cells With an Abnormal Chromosome in a Woman Previously Irradiated. Journal of Nuclear Medicine, 6, pp.740–746.

Derrien T., Estellé J., Sola SM., Knowles DG., Raineri E., Guigó R., Ribeca P. 2012. Fast computation and applications of genome mappability. PLoS ONE 7(1): e30377.

Don, P.K., Ananda, G., Chiaromonte, F. and Makova, K.D., 2013. Segmenting the human genome based on states of neutral genetic divergence. Proceedings of the National Academy of Sciences, 110(36), pp.14699-14704.

Duret, L. and Arndt, P.F., 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet, 4(5), p.e1000071.

Duret, L. & Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annual review of genomics and human genetics, 10, pp.285–311.

Durkin, S.G. and Glover, T.W., 2007. Chromosome fragile sites. Annu. Rev. Genet., 41, pp.169-192.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), pp.57-74.

Eyre-Walker, A. & Eyre-Walker, Y.C., 2014. How much of the variation in the mutation rate along the human genome can be explained? G3, 4(9), pp.1667–70.

Fitzgerald, P.H., Stewart, J. and Suckling, R.D., 1983. Retinoblastoma mutation rate in New Zealand and support for the two-hit model. Human genetics, 64(2), pp.128-130.

Flicek P., Amode MR., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fairley S., Fitzgerald S. 2011. Ensembl 2012. Nucleic Acids Research 40(Database Issue): D84-90.

Francioli, L.C. *et al.*, 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature Genetics, 46(8), pp.818-825.

Francioli LC., Polak PP., Koren A., Menelaou A., Chun S., Renkens I., van Duijn CM., Swertz M., Wijmenga C., van Ommen G., Slagboom PE., Boomsma DI., Ye K., Guryev V., Arndt PF., Kloosterman WP., de Bakker PIW., Sunyaev SR. 2015. Genome-wide patterns and properties of de novo mutations in humans. Nature Genetics 47:822–826.

Fryxell, K.J. & Moon, W.J., 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Molecular Biology and Evolution, 22(3), pp.650–658.

Gaffney, D.J. & Keightley, P.D., 2005. The scale of mutational variation in the murid genome. Genome Research, 15(8), pp.1086–1094.

Gao, G., Kasperbauer, J.L., Tombers, N.M., Wang, V., Mayer, K. and Smith, D.I., 2014. A selected group of large common fragile site genes have decreased expression in oropharyngeal squamous cell carcinomas. Genes, Chromosomes and Cancer, 53(5), pp.392-401.

Gardiner-Garden, M. & Frommer, M., 1987. CpG Islands in vertebrate genomes. Journal of Molecular Biology, 196(2), pp.261–282.

Glaab, W. & Tindall, K., 1997. Mutation rate at the hprt locus in human cancer cell lines with specific mismatch repair-gene defects. Carcinogenesis, 18(1), pp.1–8.

Glover, T.W., Berger, C., Coyle, J. and Echo, B., 1984. DNA polymerase α inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. Human genetics, 67(2), pp.136-142.

Gojobori, T., Li, W.H. & Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. Journal of Molecular Evolution, 18(5), pp.360–369.

Golding, G.B., 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol, 1(1), pp.1X-142.

Goriely, A. & Wilkie, A.O.M., 2012. Paternal Age Effect Mutations and Selfish Spermatogonial Selection: Causes and Consequences for Human Disease. The American Journal of Human Genetics, 90, pp.175–200.

Gossmann, T.I., Woolfit, M. & Eyre-Walker, A., 2011. Quantifying the variation in the effective population size within a genome. Genetics, 189(4), pp.1389–1402.

Green, P., Ewing, B., Miller, W., Thomas, P.J. and Green, E.D., 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nature genetics, 33(4), pp.514-517.

Gross, D.S. & Garrard, W.T., 1988. Nuclease hypersensitive sites in chromatin. Annual review of biochemistry, 57, pp.159–197.

Haldane, J.B.S., 1947. The Mutation Rate of the Gene for Haemophilia, and its Segregation Ratios in Males and Females. Annals of Eugenics, 13, pp.262–271.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M. and Stamatoyannopoulos, J.A., 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proceedings of the National Academy of Sciences, 107(1), pp.139-144.

Hao, D., Wang, L. & Di, L., 2016. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. Scientific Reports, 6

Harris, K., 2015. Evidence for recent, population-specific evolution of the human mutation rate. Proceedings of the National Academy of Sciences, 112(11), pp.3439-3444.

Harris, K. & Nielsen, R., 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Research, 24(9), pp.1445–1454.

Hastings, P.J., Ira, G. & Lupski, J.R., 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genetics, 5(1).

Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S. and Przeworski, M., 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. The American Journal of Human Genetics, 72(6), pp.1527-1535.

Hellmann, I., Prüfer, K., Ji, H., Zody, M.C., Pääbo, S. and Ptak, S.E., 2005. Why do human diversity levels vary at a megabase scale?. Genome research, 15(9), pp.1222-1231.

Herrero-Jimenez, P., Tomita-Mitchell, A., Furth, E.E., Morgenthaler, S. and Thilly, W.G., 2000. Population risk and physiological rate parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 447(1), pp.73-116.

Hethcote, H.W. and Knudson, A.G., 1978. Model for the incidence of embryonal cancers: application to retinoblastoma. Proceedings of the National Academy of Sciences, 75(5), pp.2453-2457.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F. and Hillman-Jackson, J., 2006. The UCSC genome browser database: update 2006. Nucleic acids research, 34(suppl 1), pp.D590-D598.

Hodgkinson, A., Chen, Y. & Eyre-Walker, A., 2012. The large-scale distribution of somatic mutations in cancer genomes. Human Mutation, 33(1), pp.136–143.

Hodgkinson, A. & Eyre-Walker, A., 2011. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics, 12(11), pp.756–766.

Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A., 2009. Cryptic variation in the human mutation rate. PLoS Biology, 7(2), pp.0226–0232.

Hollstein, M., Alexandrov, L.B., Wild, C.P., Ardin, M. and Zavadil, J., 2017. Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. Oncogene, 36(2), pp.158-167.

Hornsby, C., Page, K.M. & Tomlinson, I., 2008. The in vivo rate of somatic adenomatous polyposis coli mutation. The American journal of pathology, 172(4), pp.1062–8.

Huang FW., Hodis E., Xu MJ., Kryukov G V., Chin L., Garraway LA. 2013. Highly Recurrent TERT Promoter Mutations in Human Melanoma. Science 339:957–959.

Hughes, A.L. & Rando, O.J., 2014. Mechanisms underlying nucleosome positioning in vivo. Annual review of biophysics, 43, pp.41–63.

Hurst, L.D. & Ellegren, H., 1998. Sex biases in the mutation rate. Trends in Genetics, 14(11), pp.446–452.

Hwang, D.G. & Green, P., 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America, 101(39), pp.13994–14001.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A. and Kendall, J., 2012. De novo gene disruptions in children on the autistic spectrum. Neuron, 74(2), pp.285-299.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E. and Smith, J.D., 2014. The contribution of de novo coding mutations to autism spectrum disorder. Nature, 515(7526), pp.216-221.

Iraqui, I., Chekkal, Y., Jmari, N., Pietrobon, V., Fréon, K., Costes, A. and Lambert, S.A., 2012. Recovery of arrested replication forks by homologous recombination is error-prone. PLoS Genet, 8(10), p.e1002976.

Iwama, T., 2001. Somatic mutation rate of the APC gene. Japanese journal of clinical oncology, 31(5), pp.185–7.

Jiang, Y., Turinsky, A.L. & Brudno, M., 2015. The missing indels: An estimate of indel variation in a human genome and analysis of factors that impede detection. Nucleic Acids Research,

43(15), pp.7217–7228.
Johnson, P.L.F. & Hellmann, I., 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. Genome Biology and Evolution, 3(1), pp.842–850.
Karolchik D., Hinrichs AS., Furey TS., Roskin KM., Sugnet CW., Haussler D., Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic acids research 32:D493–D496.
Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T. and Gudjonsson, S.A., 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature, 467(7319), pp.1099-1103.
Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A. and Wong, W.S., 2012. Rate of de novo mutations and the importance of father/'s age to disease risk. Nature, 488(7412), pp.471-475.
Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R. and McCarroll, S.A., 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. The American Journal of Human Genetics, 91(6), pp.1033-1040.
Kriaucionis, S. & Bird, A., 2004. The major form of MeCP2 has a novel N-terminus generated by alternative splicing. Nucleic Acids Research, 32(5), pp.1818–1823.
Lambert, S., Watson, A., Sheedy, D.M., Martin, B. and Carr, A.M., 2005. Gross chromosomal rearrangements and elevated recombination at an inducible site-specific replication fork barrier. Cell, 121(5), pp.689-702.
Lambert, S., Mizuno, K.I., Blaisonneau, J., Martineau, S., Chanet, R., Fréon, K., Murray, J.M., Carr, A.M. and Baldacci, G., 2010. Homologous recombination restarts blocked replication forks at the expense of genome rearrangements by template exchange. Molecular cell, 39(3), pp.346-359.
Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. and Kiezun, A., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499(7457), pp.214-218.
Lee, J.A., Carvalho, C.M.B. & Lupski, J.R., 2007. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. Cell, 131(7), pp.1235–1247.
Lee, J.H., Silhavy, J.L., Kim, S., Dixon-Salazar, T., Heiberg, A., Scott, E., Bafna, V., Hill, K.J., Collazo, A., Funari, V. and Russ, C., 2012. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nature genetics, 44(8), pp.941-945.
Lercher, M.J. & Hurst, L.D., 2002. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? Gene, 300(1–2), pp.53–58.
Lercher, M.J. & Hurst, L.D., 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends in Genetics, 18(7), pp.337–340.
Lercher, M.J., Williams, E.J.B. & Hurst, L.D., 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. Molecular Biology and Evolution, 18(11), pp.2032–2039.
Le Tallec, B., Dutrillaux, B., Lachages, A.M., Millot, G.A., Brison, O. and Debatisse, M., 2011. Molecular profiling of common fragile sites in human fibroblasts. Nature structural & molecular biology, 18(12), pp.1421-1423.
Le Tallec, B., Millot, G.A., Blin, M.E., Brison, O., Dutrillaux, B. and Debatisse, M., 2013. Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. Cell reports, 4(3), pp.420-428.
Letessier, A., Millot, G.A., Koundrioukoff, S., Lachagès, A.M., Vogt, N., Hansen, R.S., Malfoy, B., Brison, O. and Debatisse, M., 2011. Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature, 470(7332), pp.120-123.

Lichtenauer-Kaligis, E.G., Thijssen, J., den Dulk, H., van de Putte, P., Tasseron-de Jong, J.G. and Giphart-Gassler, M., 1996. Comparison of spontaneous hprt mutation spectra at the nucleotide sequence level in the endogenous hprt gene and five other genomic positions. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 351(2), pp.147-155.

Liu, L., De, S. & Michor, F., 2013. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nature communications, 4, p.1502. Available at: http://dx.doi.org/10.1038/ncomms2502.

Liu, P., Carvalho, C.M., Hastings, P.J. and Lupski, J.R., 2012. Mechanisms for recurrent and complex human genomic rearrangements. Current opinion in genetics & development, 22(3), pp.211-220.

Löytynoja, A., Goldman, N., Campus, W.G. and Hinxton, U.K., 2016. Short template switch events in human evolution cause complex mutation patterns. bioRxiv, p.038380.

Luebeck, E.G. & Moolgavkar, S.H., 2003. Multistage Carcinogenesis and the Incidence of Human Cancer. Genes Chromosomes and Cancer, 38(4), pp.302–306.

Lu, J., Li, H., Hu, M., Sasaki, T., Baccei, A., Gilbert, D.M., Liu, J.S., Collins, J.J. and Lerou, P.H., 2014. The distribution of genomic variations in human iPSCs is related to replication-timing reorganization during reprogramming. Cell reports, 7(1), pp.70-78.

Lukusa, T. & Fryns, J.P., 2008. Human chromosome fragility. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms, 1779(1), pp.3–16.

Lynch, M., 2010. Evolution of the mutation rate. Trends in Genetics 26:345–352.

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America, 107(3), pp.961–8.

Makova, K.D. and Hardison, R.C., 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nature Reviews Genetics, 16(4), pp.213-223.

Makova, K.D. and Li, W.H., 2002. Strong male-driven evolution of DNA sequences in humans and apes. Nature, 416(6881), pp.624-626.

Mannucci, P.M. and Tuddenham, E.G., 2001. The hemophilias—from royal genes to gene therapy. New England Journal of Medicine, 344(23), pp.1773-1779.

Margoliash, E., 1963. Primary Structure and evolution of cytochrome c. Proceedings of the National Academy of Sciences of the United States of America, 50(1961), pp.672–679.

Martinocorena, I. & Campbell, P.J., 2015. Somatic mutation in cancer and normal cells. Science, 349(6255), pp.1483–1489.

Matassi, G., Sharp, P.M. & Gautier, C., 1999. Chromosomal location effects on gene sequence evolution in mammals. Current Biology, 9(15), pp.786–791.

McVean, G.T. & Hurst, L.D., 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature, 386(6623), pp.388–92.

McVicker, G., Gordon, D., Davis, C. and Green, P., 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet, 5(5), p.e1000471.

Melis, J.P.M., van Steeg, H. & Luijten, M., 2013. Oxidative DNA damage and nucleotide excision repair. Antioxidants & redox signaling, 18(18), pp.2409–19.

Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A. and Wu, W., 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell, 151(7), pp.1431-1442.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S. and Devine, S.E., 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome research, 16(9), pp.1182-1190.

Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., Webber, C. and Devine, S.E., 2011. Natural genetic variation caused by small insertions and deletions in the human genome. Genome research, pp.gr-115907.

Minoche AE., Dohm JC., Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology 12:R112.

Mizuno, K. *et al.*, 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. Genes and Development, 23(24), pp.2876–2886.

Mizuno, K.I., Miyabe, I., Schalbetter, S.A., Carr, A.M. and Murray, J.M., 2013. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. Nature, 493(7431), pp.246-249.

Mizunol, S., Watanabe, S., & Iwama, T., 1993. Colorectal Cancer Incidence Linking Model in Familial Adenomatous Polyposis and the General Population. , (c), pp.109–115.

Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A.M., Brinkman, A.B., Martin, S., Ramakrishna, M. and Butler, A., 2016. The topography of mutational processes in breast cancer genomes. Nature communications, 7.

Mrasek, K., Schoder, C., Teichmann, A.C., Behr, K., Franze, B., Wilhelm, K., Blaurock, N., Claussen, U., Liehr, T. and Weise, A., 2010. Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. International journal of oncology, 36(4), p.929.

Mugal, C.F., von Grunberg, H.-H. & Peifer, M., 2009. Transcription-Induced Mutational Strand Bias and Its Effect on Substitution Rates in Human Genes. Mol Biol Evol, 26(1), pp.131–142.

Mullaney, J.M. *et al.*, 2010. Small insertions and deletions (INDELs) in human genomes. Human Molecular Genetics, 19(R2), pp.131–136.

Nachman, M.W. & Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics, 156(1), pp.297–304.

Nazarian R., Shi H., Wang Q., Kong X., Koya RC., Lee H., Chen Z., Lee M-K., Attar N., Sazegar H., Chodon T., Nelson SF., McArthur G., Sosman JA., Ribas A., Lo RS. 2010. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. Nature 468:973–7.

Neale, B.M., Kou, Y., Liu, L., Ma'Ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V. and Polak, P., 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature, 485(7397), pp.242-245.

Nelder JA., Mead R., Nelder BJA., Mead R. 1965. A Simplex Method for Function Minimization. The Computer Journal 7:308–313.

Ness, R.W., Morgan, A.D., Vasanthakrishnan, R.B., Colegrave, N. and Keightley, P.D., 2015. Extensive de novo mutation rate variation between individuals and across the genome of Chlamydomonas reinhardtii. Genome research, 25(11), pp.1739-1749.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M. and Shlien, A., 2012. The life history of 21 breast cancers. Cell, 149(5), pp.994-1007.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C. and Van Loo, P., 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature, 534(7605), pp.47-54.

O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E. and Snyder, M.P., 2012. Extensive genetic variation in somatic human tissues. Proceedings of the National Academy of Sciences, 109(44), pp.18018-18023.

O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., MacKenzie, A.P., Ng, S.B., Baker, C. and Rieder, M.J., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics, 43(6), pp.585-589.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C.,

Smith, J.D. and Turner, E.H., 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature, 485(7397), pp.246-250.

Ozeri-Galai, E., Bester, A.C. & Kerem, B., 2012. The complex basis underlying common fragile site instability in cancer. Trends in Genetics, 28(6), pp.295–302.

Panchin, A.Y., Makeev, V.J. & Medvedeva, Y.A., 2016. Preservation of methylated CpG dinucleotides in human CpG islands. Biology direct, 11(1), p.11.

Paten, B., Herrero, J., Beal, K., Fitzgerald, S. and Birney, E., 2008. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome research, 18(11), pp.1814-1828.

Paten, B., Herrero, J., Beal, K. and Birney, E., 2009. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics, 25(3), pp.295-301.

Petermann, E., Orta, M.L., Issaeva, N., Schultz, N. and Helleday, T., 2010. Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-mediated pathways for restart and repair. Molecular cell, 37(4), pp.492-502.

Pink, C.J. & Hurst, L.D., 2010. Timing of replication is a determinant of neutral substitution rates but does not explain slow y chromosome evolution in rodents. Molecular Biology and Evolution, 27(5), pp.1077–1086.

Pinto, Y., Gabay, O., Arbiza, L., Sams, A.J., Keinan, A. and Levanon, E.Y., 2016. Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity. Genome research, 26(5), pp.579-587.

Poduri, A., Evrony, G.D., Cai, X., Elhosary, P.C., Beroukhim, R., Lehtinen, M.K., Hills, L.B., Heinzen, E.L., Hill, A., Hill, R.S. and Barry, B.J., 2012. Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron, 74(1), pp.41-48.

Poduri, A., Evrony, G.D., Cai, X. and Walsh, C.A., 2013. Somatic mutation, genomic variation, and neurological disease. Science, 341(6141), p.1237758.

Polak, P. & Arndt, P.F., 2008. Transcription induces strand-specific mutations at the 5??? end of human genes. Genome Research, 18(8), pp.1216–1223.

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M.S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J.A. and Sunyaev, S.R., 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature, 518(7539), pp.360-364.

Prendergast, J.G.D. *et al.*, 2007. Chromatin structure and evolution in the human genome. BMC evolutionary biology, 7, p.72.

Price, E.A., Price, K., Kolkiewicz, K., Hack, S., Reddy, M.A., Hungerford, J.L., Kingston, J.E. and Onadim, Z., 2013. Spectrum of RB1 mutations identified in 403 retinoblastoma patients. Journal of medical genetics, pp.jmedgenet-2013.

Quail, M.A., Kozarewa I., Smith F., Scally A., Stephens P.J., Durbin R., Swerdlow H., Turner D.J. 2008. A large genome center's improvements to the Illumina sequencing system. Nature methods 5:1005–10.

R Core Team., 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Rivière, J.B., Mirzaa, G.M., O'Roak, B.J., Beddaoui, M., Alcantara, D., Conway, R.L., St-Onge, J., Schwartzentruber, J.A., Gripp, K.W., Nikkel, S.M. and Worthylake, T., 2012. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nature genetics, 44(8), pp.934-940.

Roberts, S.A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L.J., Kryukov, G.V., Malc, E., Mieczkowski, P.A. and Resnick, M.A., 2012. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Molecular cell, 46(4), pp.424-435.

Roberts, S.A. & Gordenin, D.A., 2014. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer, 14(12), pp.786–800.

Rosenfeld, J.A., Malhotra, A.K. & Lencz, T., 2010. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. Nucleic Acids Research, 38(18), pp.6102–6111.

Sakofsky, C.J., Ayyar, S., Deem, A.K., Chung, W.H., Ira, G. and Malkova, A., 2015. Translesion polymerases drive microhomology-mediated break-induced replication leading to complex chromosomal rearrangements. Molecular cell, 60(6), pp.860-872.

Sánchez-Sánchez, F., Ramírez-Castillejo, C., Weekes, D.B., Beneyto, M., Prieto, F., Nájera, C. and Mittnacht, S., 2007. Attenuation of disease phenotype through alternative translation initiation in low-penetrance retinoblastoma. Human mutation, 28(2), p.159.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L. and Walker, M.F., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature, 485(7397), pp.237-241.

Schrider, D.R., Houle, D., Lynch, M. and Hahn, M.W., 2013. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. Genetics, 194(4), pp.937-954.

Schrider, D.R., Hourmozdi, J.N. & Hahn, M.W., 2011. Pervasive multinucleotide mutational events in eukaryotes. Current Biology, 21(12), pp.1051–1054. Available at: http://dx.doi.org/10.1016/j.cub.2011.05.013.

Schuster-Bockler, B. & Lehner, B., 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature, 488(7412), pp.504–507.

Seplyarskiy, V.B., Andrianova, M.A. & Bazykin, G.A., 2016. APOBEC3A / B-induced mutagenesis is responsible for 20 % of heritable mutations in the TpCpW context.

Seplyarskiy, V.B., Bazykin, G.A. & Soldatov, R.A., 2014. Polymerase ζ activity is linked to replication timing in humans : evidence from mutational signatures.

Sigurðardóttir, S., Helgason, A., Gulcher, J.R., Stefansson, K. and Donnelly, P., 2000. The mutation rate in the human mtDNA control region. The American Journal of Human Genetics, 66(5), pp.1599-1609.

Smith, D.I., Zhu, Y., McAvoy, S. and Kuhn, R., 2006. Common fragile sites, extremely large genes, neural development and cancer. Cancer letters, 232(1), pp.48-57.

Smith, N.G.C. & Lercher, M.J., 2002. Regional similarities in polymorphism in the human genome extend over many megabases. Trends in Genetics, 18(6), pp.281–283.

Smith, N.G.C., Webster, M.T. & Ellegren, H., 2003. A low rate of simultaneous double-nucleotide mutations in primates. Molecular Biology and Evolution, 20(1), pp.47–53.

Smith, T., Ho, G., Christodoulou, J., Price, E.A., Onadim, Z., Gauthier-Villars, M., Dehainault, C., Houdayer, C., Parfait, B., Minkelen, R. and Lohman, D., 2016. Extensive variation in the mutation rate between and within human genes associated with Mendelian disease. Human mutation, 2016 Feb 1.

Spencer, C.C.A. *et al.*, 2006. The influence of recombination on human genetic diversity. PLoS Genetics, 2(9), pp.1375–1385.

Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M. and Sunyaev, S.R., 2009. Human mutation rate associated with DNA replication timing. Nature genetics, 41(4), pp.393-395.

Stone, J.E., Lujan, S.A. and Kunkel, T.A., 2012. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in Saccharomyces cerevisiae. Environmental and molecular mutagenesis, 53(9), pp.777-786.

Stratton, M., Campbell, P. & Futreal, P., 2009. The cancer genome. Nature, 458(7239), pp.719–724. Available at: http://www.nature.com/nature/journal/v458/n7239/abs/nature07943.html.

Subramanian, S. & Kumar, S., 2003. Neutral substitutions occur at a faster rate in exons than in

noncoding DNA in primate genomes. Genome Research, 13(5), pp.838–844.
Supek, F. & Lehner, B., 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature, 521(7550), pp.81–84.
Sved, J. & Bird, a, 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proceedings of the National Academy of Sciences of the United States of America, 87(12), pp.4692–4696.
Terekhanova, N.V., Bazykin, G.A., Neverov, A., Kondrashov, A.S. and Seplyarskiy, V.B., 2013. Prevalence of multinucleotide replacements in evolution of primates and Drosophila. Molecular biology and evolution, 30(6), pp.1315-1325.
The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature, 526(7571), pp.68–74.
The Chimpanzee Sequencing; A Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 437(7055), pp.69–87.
Thurman, R.E., Day, N., Noble, W.S. and Stamatoyannopoulos, J.A., 2007. Identification of higher-order functional domains in the human ENCODE regions. Genome research, 17(6), pp.917-927.
Tomso, D.J. & Bell, D.A., 2003. Sequence context at human single nucleotide polymorphisms: Overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. Journal of Molecular Biology, 327(2), pp.303–308.
Treangen, T.J., Salzberg, S.L. 2013. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13:36–46.
Tyekucheva, S., Makova, K.D., Karro, J.E., Hardison, R.C., Miller, W. and Chiaromonte, F., 2008. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome biology, 9(4), p.R76.
Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J. and Yagi, T., 2015. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome research, 25(8), pp.1125-1134.
Varki, A. & Altheide, T.K., 2005. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. Genome Research, 15(12), pp.1746–1758.
Veltman, J. a & Brunner, H.G., 2012. De novo mutations in human genetic disease. Nature reviews. Genetics, 13(8), pp.565–75.
Walsh, E. *et al.*, 2013. Mechanism of replicative DNA polymerase delta pausing and a potential role for DNA polymerase kappa in common fragile site replication. Journal of Molecular Biology, 425(2), pp.232–243.
Whitlock, M.C., 2005. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. Journal of Evolutionary Biology, 18(5), pp.1368–1373.
Williams, E.J. & Hurst, L.D., 2000. The proteins of linked genes evolve at similar rates. Nature, 407(6806), pp.900–903.
Wilson, T.E., Arlt, M.F., Park, S.H., Rajendran, S., Paulsen, M., Ljungman, M. and Glover, T.W., 2015. Large transcription units unify copy number variants and common fragile sites arising under replication stress. Genome research, 25(2), pp.189-200.
Wolfe, K.H., Sharp, P.M. & Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. Nature, 337(6204), pp.283–5. Available at: http://dx.doi.org/10.1038/337283a0.
Wong, W.S., Solomon, B.D., Bodian, D.L., Kothiyal, P., Eley, G., Huddleston, K.C., Baker, R., Thach, D.C., Iyer, R.K., Vockley, J.G. and Niederhuber, J.E., 2016. New observations on maternal age effect on germline de novo mutations. Nature communications, 7.
Woo, Y.H. & Li, W.-H., 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nature Communications, 3, p.1004.
Yadav, V.K., Degregori, J. & De, S., 2016. The landscape of somatic mutations in protein coding

genes in apparently benign human tissues carries signatures of relaxed purifying selection. Nucleic Acids Research, 44(5), pp.2075–2084.

Yaffe, E., Farkash-Amar, S., Polten, A., Yakhini, Z., Tanay, A. and Simon, I., 2010. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet, 6(7), p.e1001011.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular evolution, 39(3), pp.306-314.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. and Girón, C.G., 2015. Ensembl 2016. Nucleic acids research, p.gkv1157.

Ying, H., Epps, J., Williams, R. and Huttley, G., 2010. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. Molecular biology and evolution, 27(3), pp.637-649.

Yurov, Y.B., 2005. The Variation of Aneuploidy Frequency in the Developing and Adult Human Brain Revealed by an Interphase FISH Study. Journal of Histochemistry and Cytochemistry, 53(3), pp.385–390.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B. and Wong-Erasmus, M., 2011. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database, 2011, p.bar026.

Zhao, Z. & Boerwinkle, E., 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome. Genome Research, 12(11), pp.1679–1686.

Zhuang J., Wang J., Theurkauf W., Weng Z. 2014. TEMP: A computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Research 42:6826–6838.

Zuckerkandl, E. & Pauling, L., 1965. Evolutionary divergence and convergence in proteins. Evolving Genes and Proteins, pp.97–166.

Appendices

Appendix 2.1

				Number of sites with x DNMs																	
Gene	**Mutational category**	**Number of sites**	**Number of DNMs**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**52**	**54**	**61**	**85**
RB1	CTS	11	97	0	0	0	0	0	1	1	2	1	2	1	1	1	1				
	CTV	9	5	6	2	0	1														
	NTS	44	15	33	8	2	1														
	NTV	327	32	301	20	6															
NF1	CTS	18	52	1	3	5	6	1	0	0	1	0	0	1							
	CTV	12	4	10	0	2															
	NTS	187	24	166	18	3															
	NTV	851	20	831	20																
MECP2	CTS	5	252	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
	CTV	7	0	7																	
	NTS	16	6	12	2	2															
	NTV	152	12	145	6	0	0	0	0	1											
MECP2 - restricted	CTS	4	252	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
	CTV	3	0	3																	
	NTS	5	2	4	0	1															
	NTV	38	10	33	4	0	0	0	0	1											

Appendix 2.1. The complete dataset. Given are the number of sites with x DNMs, including those sites with no DNMs, but which when mutated would give a nonsense mutation.

Appendix 2.2

Hits	Houdayer	Lohmann	Onadim
0	0	0	2
1	1	1	4
2	2	2	2
3	2	4	1
4	4	1	0
5	2	2	1
6	0	0	1
7	0	1	0
Probability	0.85	0.81	0.086

Appendix 2.2. The number of sites hit by a CpG transition nonsense DNM in RB1 in the three datasets, along with the probability from a test of heterogeneity.

Appendix 2.3

Site	Houdayer	Lohmann	Onadim
751	4	3	0
763	1	3	1
958	5	4	0
1072	4	3	6
1333	4	7	1
1363	3	5	3
1399	5	2	2
1654	3	2	2
1666	2	1	5
1735	2	3	1
2359	4	5	1
Total DNMs	37	38	22

Appendix 2.3. he number of DNMs at each site in RB1 at which a CpG transition causes a nonsense mutation in the three datasets. The three datasets do not show a significant difference in the pattern of mutation as assessed by a chi-square test, generating the null distribution by randomization.

Appendix 2.4

Model no.	Model description	N	Log liklihood
1	Gamma distribution shared across genes and mutational categories	12	-101.39
2	**Separate gamma distribution for each mutational category shared across genes**	**15**	**-86.91**
3	Separate gamma distribution for each gene shared across mutational categories	14	-98.89
4	Separate gamma distribution for each gene and mutational category combination	22	-83.56

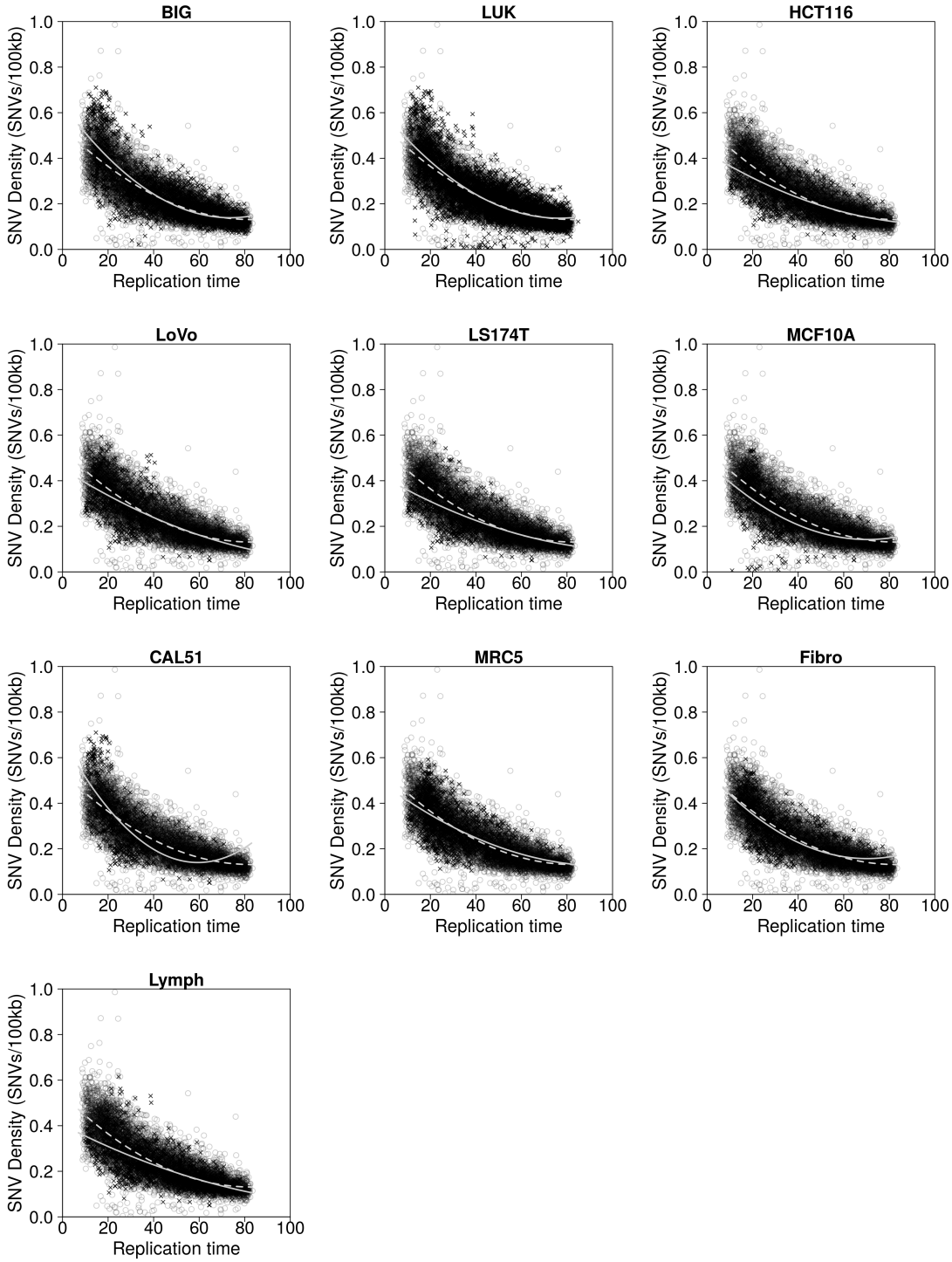
Appendix 2.4. Log likelihood values for various models. The analysis is performed on the restricted MECP2 data. N is the number of parameters in each model. The figures in bold indicate the model favored by likelihood ratio tests - model 2 provides a significantly better fit to the data than model 1, whereas model 3 does not, and model 4 does not provide a significantly better fit than model 2.

Appendix 3.1

GenomicPosition	RIKEN	NCC	sum
X:56209339	6	0	6
10:96652829	6	0	6
10:96652827	6	0	6
X:56209340	5	0	5
5:85091859	5	0	5
5:1295228	0	5	5
9:121267366	4	0	4
8:119547627	4	0	4
19:22314552	1	2	3
14:95832895	1	2	3
9:16932821	2	1	3
7:27901228	2	1	3
4:162437670	2	1	3
3:164903710	2	1	3
Y:4796240	3	0	3
X:84996701	3	0	3
7:11432162	3	0	3
7:11432157	3	0	3
3:174306603	3	0	3
2:49173787	3	0	3
2:139556678	3	0	3
19:8673262	3	0	3
1:190881448	3	0	3
X:79125571	0	3	3
6:78532352	0	3	3
5:97912191	0	3	3
4:190837614	0	3	3
19:44959650	0	3	3
15:73206445	0	3	3
14:74659965	0	3	3

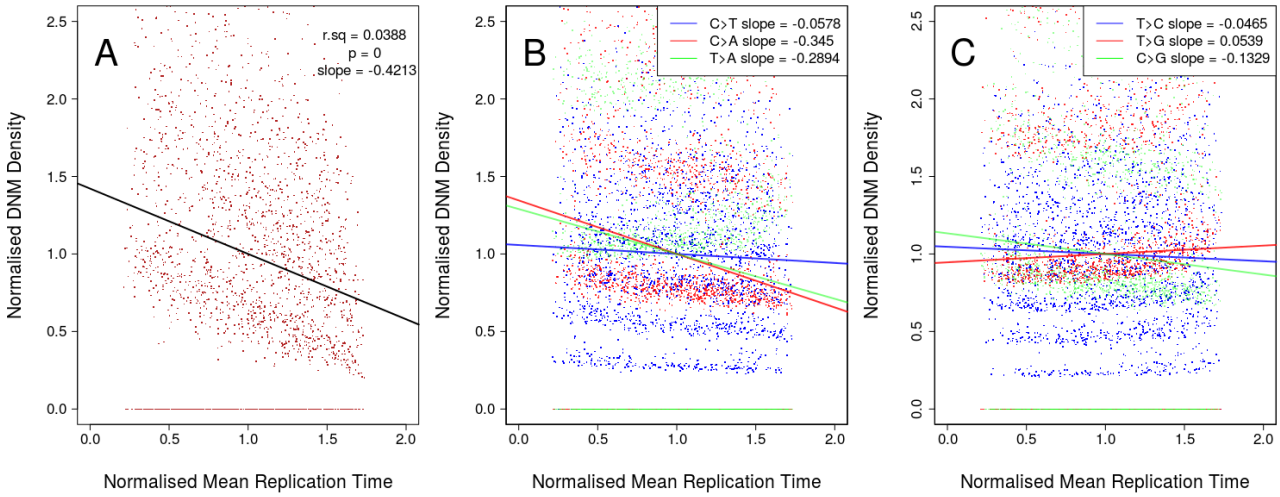
Appendix 3.1. Genomic positions and seqeuncing centre origin of the excess liver cancers used for privacy calculations. NCC = National Cancer Centre Japan, RIKEN = RIKEN institute Japan.

Appendix 4.1



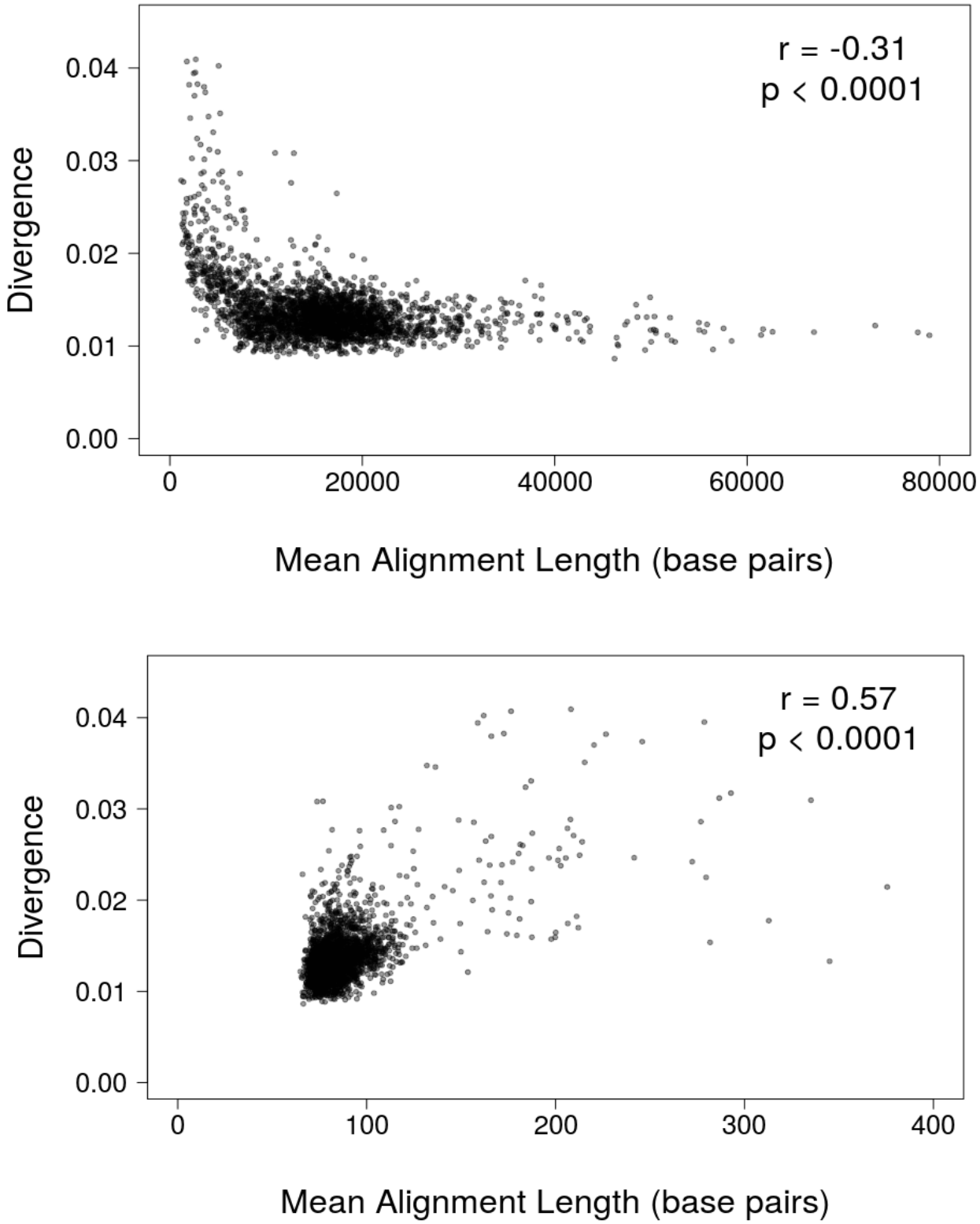
Appendix 4.1. The fit of separate quadratic terms for each aCFS tissue type and study (solid lineand crosses) and non-FS (dashed line and circles) windows.

Appendix 5.1



Appendix 5.1. Graphs showing DNM density as a product of replication time. each point represents one 1Mb window, values for both axes were normalised by dividing the score for each window by the mean for all windows, A, CpG transitions, B, strong (C/G) to weak (T/A) and weak to weak mutations, C, weak to strong and strong to strong mutations.

Appendix 5.2



Appendix 5.2. Graphs showing mean alignment lengths versus divergence for the PW alignments (top panel) and the MZ alignments (bottom panel). Spearman's correlation coefficient was used to obtain r.

Appendix 7. The variable association of replication time with SNV density.

The following observations arose tangentally to Chapter 4 whilst investigating the relationship of replication time with fragile sites and SNV density. I felt that this work was important but did not fit within the framework of this thesis nor was it substantial enough to stand alone as a chapter. Therefore I have decided to include it in the appendix as a short aside highlighting the variability of the association of replication time with SNV density in cancers.

The effect of replication time on mutation has been widely documented in both germ-line and somatic tissues and it has become widely accepted that later replicating DNA harbours an increased rate of point mutations (Stamatoyannopoulos *et al.*, 2009; Chen *et al.*, 2010; Koren *et al.*, 2012). We refer to this as phenomenon as replication time dependent mutagenesis (RTAM). In somatic tissues it has been claimed that functional mismatch repair (MMR) underlies RTAM (Supek & Lehner, 2015). We show that, in a cohort of 160 samples that there is extensive variation in RTAM that is not linked to MMR deficiency and demonstrate that other determinants of RTAM are at least as important in cancer samples.

From the 507 publicly available whole genome sequenced cancer samples published by Alexandrov *et al.* (Alexandrov *et al.*, 2013), we selected 160 samples which contained over 3,000 SNVs each. These were comprised of 24 lung, 78 liver, 10 pancreatic and 48 breast cancers, none of which are commonly associated with MMR deficiency (Cunningham *et al.*, 1998; Miyakura *et al.*, 2001;

Chiappini *et al.*, 2004; Gargiulo *et al.*, 2009). We examined the distribution of SNVs in these sample as a function of replication time. Using the mean replication time of 5 cell lines (Hansen *et al.*, 2010) (see methods) we divided the genome into quintiles. We quantified the level of RTAM in terms of a quintile enrichment score (QES), which is the number of SNVs in the latest replicating quintile divided by the number of SNVs in earliest replicating quintile; we favoured this method as opposed to a linear regression as there is evidence that there is a non-linear relationship in the extremely late replicating regions. QES differs significantly amongst cancer genomes ranging 6-fold from 0.81 to 5.42 (Chi-square test $p < 10^{-16}$) (Fig. 1, also see specific examples in Fig. 2a-f). Unfortunately, not being in possession of the original 160 samples, we could not directly test them for MMR deficiency by looking for micro satellite instability (MSI) in the usual manner (Umar *et al.*, 2004). However, mutational signatures 6, 15, 20 and 26 from Alexandrov *et al.* (Alexandrov *et al.*, 2013) are indicative of MMR. To date, signature 15 has never been reported in the 4 cancer types covered in these 160 samples. Signatures 20 and 26 have been found in breast cancers, but only at frequencies <3%, and signature 6 has been found across all cancers, but again only at low frequencies (<3%) (http://cancer.sanger.ac.uk/cosmic/signatures). From the 160 samples, Only one sample, HX13T, showed sufficient levels of signature 6 to be classified as MMR deficient. HX13T does exhibit an attenuated QES of 1.27 (Fig. 2c). A further 10 samples (Supplemental data) exhibit very low levels of signature 6 that could be due to low-level sub-clonal MSI or limitations of the methodology used to detect the signatures, however the remaining 149 samples contain no traces of any of the MMR signatures (Alexandrov, L., personal communication). Furthermore, none of these samples contained exonic mutations in *MLH1*, *MSH2*, *MSH6* and *PMS2*, mutations in which genes cause MSI in >90% of Lynch syndrome cases (Fishel *et al.*, 1993; Bronner *et al.*, 1994).

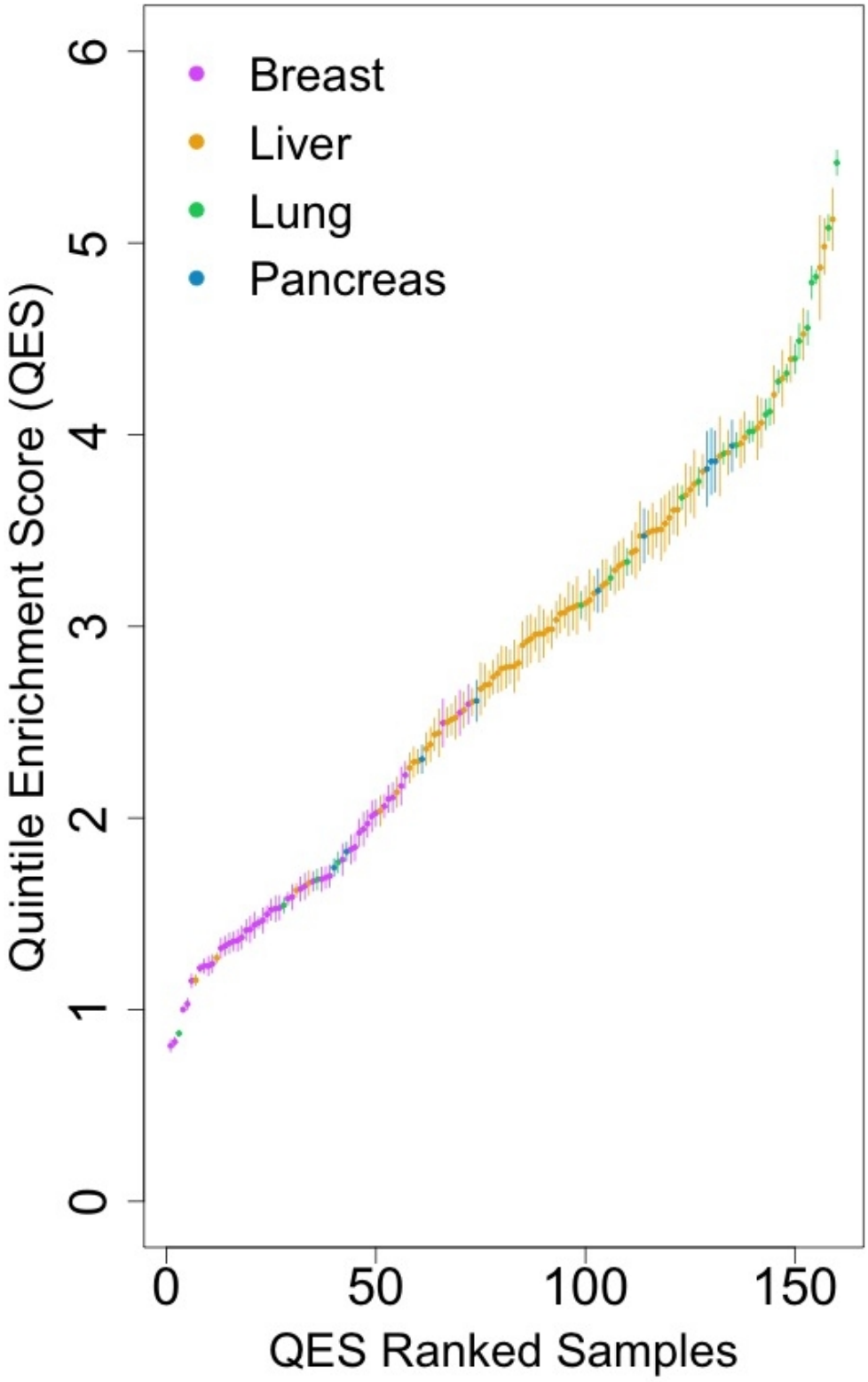


Figure 1. 160 samples ranked by their quintile enrichment scores and coloured by cancer type. Vertical lines represent 1 standard deviation

We identified 4 samples (LUAD-S01345, PD4120a, PD4199a and PD6722a) from the 149 MMR proficient samples that demonstrated a complete lack of RTAM (Fig. 2a-b) (i.e. QES ≤ 1). The low QES found in these 4 samples is in stark contrast to a QES of ~2-6 that was found in previous studies in humans (Koren *et al.*, 2012) and yeast (Lang & Murray, 2011) and the median QES of 2.78 found when considering all 160 samples (Fig. 2e). It was conspicuous that 3 of the 4 low-QES samples were breast cancers. Ranking each of the 160 samples by their QES showed that breast cancers dominated the low scores (Fig. 1 and Supplemental data) and confirmed that there are significant differences between cancer types (Kruskel-Wallis test $p < 10^{-16}$) (Fig 2f). Strikingly the median QES for breast cancers is 1.56, only slightly greater than the median QES amongst the MMR-deficient cancers as studied by Supek and Lehner (median QES scores for MMR deficient colorectal cancer = 1.47, uterine cancer = 1.30, stomach cancer = 1.43 and pooled tissues = 1.42, calculated from original data in [1] provided by Supek & Lehner). To exclude cell line specificity of the replication timing data being the cause of the low median QES for breast cancers, we also used replication time data from the breast cancer derived MCF-7 cell line (Hansen *et al.*, 2010). In this case similar results were obtained (data not shown). Additionally it has been shown that regions of the genome exhibiting extremely late/early replication time are remarkably stable [REF]. It is therefore evident that not only is there very large variation in the degree of RTAM that is not associated with MMR status, but that cancers can have no RTAM even when they are MMR proficient.

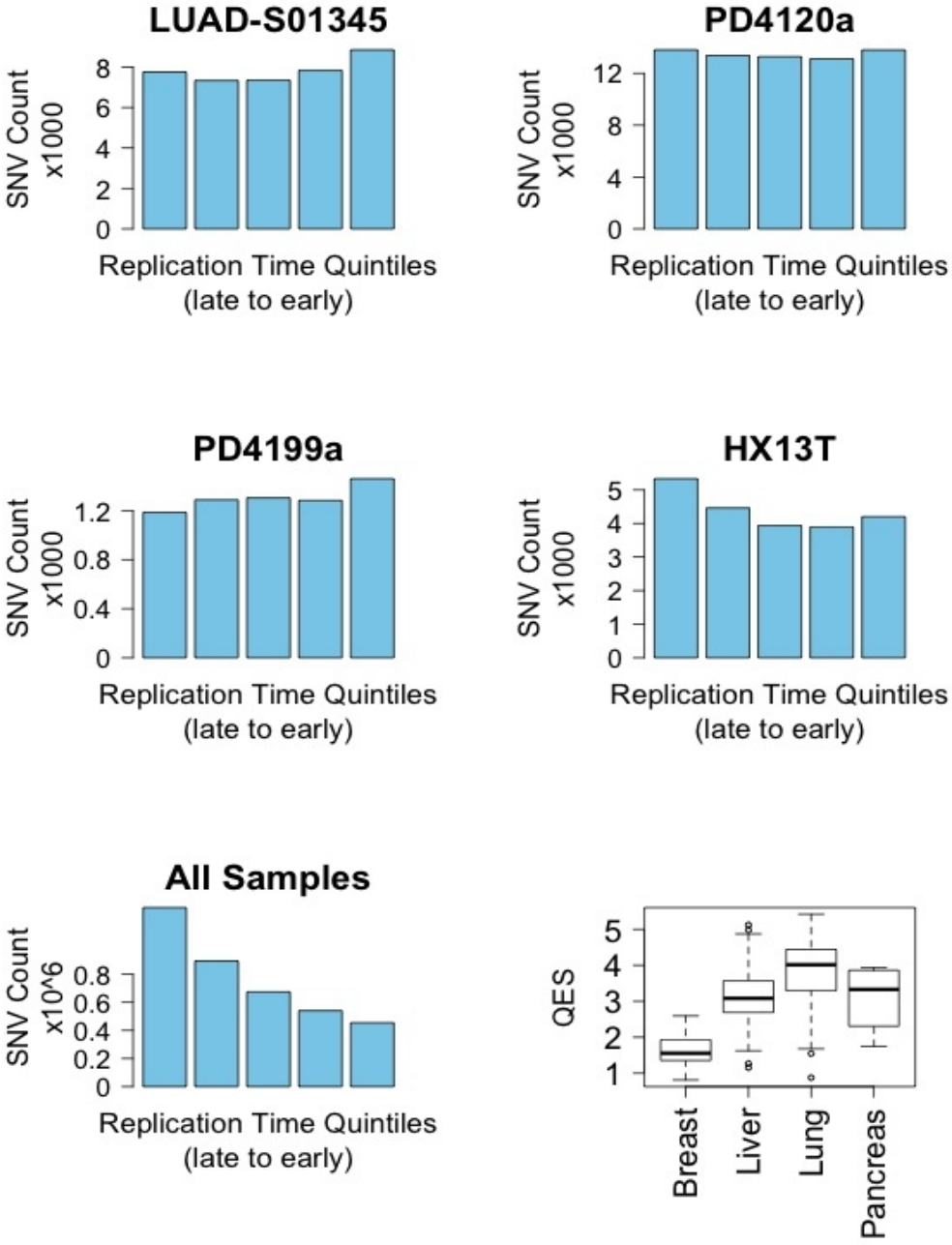


Figure 2. Variation of replication time associated mutagenesis (RTAM) across mismatch repair (MMR) proficient cancer samples. **a-c**, Single nucleotide variant (SNV) counts across replication time quintiles for 3 low-QES samples, LUAD-S01345, PD4120a, PD4199a, which exhibit a lack of RTAM. **d**, Same as a-c but for HX13T, the only confirmed MSI sample **e**, Same as a-d but summed across all 160 samples illustrating the median QES = 2.78. **f**, Boxplot of median QES for each cancer type, box limits indicate inter quartile ranges (IQR), whiskers denote the limits of extreme values or 1.5xIQR, whichever is lesser, open circles denote outlying samples.

The reason why breast cancers show low RTAM is unclear. However it is interesting to note that the 4 low-QES samples all show high levels of signatures 2 and 13, indicative of of APOBEC induced hyper-mutation (Alexandrov, L., personal communication). As estrogen has been shown to induce expression of the AID/APOBEC family of deaminases (Pauklin *et al.*, 2009), one hypothesis could be that hormonally driven cancers are more likely to over-express APOBEC proteins, and the resulting hyper-mutation may occur independently of replication time. Since the initial discovery of this pattern, it has been shown that APBOEC over-expression is a likely candidate for the the lack of RTAM (Morganella *et al.*, 2016).

We conclude that although MMR may be a determinant of RTAM in cancers that are commonly associated with MMR, such as colorectal, stomach or uterine cancers, it is rarely likely to be the major determinant of RTAM in other cancers. Not only have our findings shown the existence of MMR proficient cancer samples with a complete lack of RTAM, they have also shown that the extent of RTAM varies >6 fold across diverse cancer types, the vast majority of which are not MMR deficient and that >10% of cancer samples exhibit attenuated RTAM to a level lower than the MMR deficient median. We also show that RTAM is associated with cancer type, with relatively low QES in breast cancers and the highest QES in lung cancers (Fig 1, Fig 2f). Therefore upon this evidence, we suggest that other determinants, including, but not limited to, cancer type and APOBEC over-expression, likely contribute to RTAM to a greater degree than does variable MMR.

Methods

Replication timing data and dividing the genome into 100 kb windows.

We downloaded Encode Repli-seq wavelet smoothed signal data (Hansen *et al.*, 2010) aligned to the reference genome (HG19/GRCh37) for the GM12878, HeLa, HUVEC, K562, HepG2 and MCF-7

cell lines from the UCSC ftp site ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/. We computed the mean replication time for chromosomes 1-22 and X in non-overlapping 100 kb windows across the GM12878, HeLa, HUVEC, K562, HepG2 cell lines. We used the Repli-Seq start coordinates for each chromosome (bp 24500) to denote the beginning of the first window. Chromosome Y was not included as no replication time data was available. Any windows containing unknown bases (denoted as N in the reference genome) were excluded from the analysis. This resulted in 28,080 100 kb windows remaining. The windows were ranked by their mean replication times and divided into quintiles each containing 5616 100 kb windows. We also used replication time data from the MCF-7 cell line to control for tissue specific effects.

Single nucleotide variants.

The 160 Sample IDs were extracted from Supplementary Table 1 in (Supek & Lehner, 2015). Single nucleotide variants (SNVs) for breast, liver, lung and pancreatic cancers from (Alexandrov *et al.*, 2013) were downloaded from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/mutational_catalogs/genomes/ and filtered to retain only the SNVs from the 160 samples resulting in a total of 2,838,468 SNVs. These SNVs were assigned to their relevant 100 kb windows and total SNV and total counts per window was calculated. Each window was assigned to its relevant replication time quintile and SNVs summed for each quintile.

Detection of MMR deficiency in samples.

MMR status of 160 samples was confirmed via personal communication from Ludmil Alexandrov after determination of levels of 4 known MMR signatures 6,16,20 and 26 from (Alexandrov *et al.*, 2013).

Statistical analysis.

Quintile enrichment scores were calculated by dividing the sum of SNVs in the latest replicating quintile (Q1) by sum of SNVs in the earliest replicating quintile (Q1). R version 3.1.3 was used for all statistics (stats package); Chi Squared test was used to test for significant differences in QES between the 160 samples, Kruskel-Wallis test was used to test for significant differences in the median QES between cancer types and variances and standard deviations were calculated for each QES shown in figure 2. QES scores for the MSI colorectal, stomach, uterine and pooled intergenic regions were manually calculated from the original data used to produce the graphs 2e-g in (Supek & Lehner, 2015) by dividing the median relative SNV frequency in the latest replicating quintile by the median SNV frequency in the earliest replicating quintile.

References.

Alexandrov LB., Nik-Zainal S., Wedge DC., Aparicio S a JR., Behjati S., Biankin A V., Bignell GR., Bolli N., Borg A., Børresen-Dale A-L., Boyault S., Burkhardt B., Butler AP., Caldas C., Davies HR., Desmedt C., Eils R., Eyfjörd JE., Foekens J a., Greaves M., Hosoda F., Hutter B., Ilicic T., Imbeaud S., Imielinski M., Imielinsk M., Jäger N., Jones DTW., Jones D., Knappskog S., Kool M., Lakhani SR., López-Otín C., Martin S., Munshi NC., Nakamura H., Northcott P a., Pajic M., Papaemmanuil E., Paradiso A., Pearson J V., Puente XS., Raine K., Ramakrishna M., Richardson AL., Richter J., Rosenstiel P., Schlesner M., Schumacher TN., Span PN., Teague JW., Totoki Y., Tutt ANJ., Valdés-Mas R., van Buuren MM., van 't Veer L., Vincent-Salomon A., Waddell N., Yates LR., Zucman-Rossi J., Futreal PA., McDermott U., Lichter P., Meyerson M., Grimmond SM., Siebert R., Campo E., Shibata T., Pfister SM., Campbell PJ., Stratton MR. 2013. Signatures of mutational processes in human cancer. *Nature* 500:415–21. DOI: 10.1038/nature12477.

Bronner C., Baker S., Morrison P., Warren G., Smith L., Lescoe M., Kane M., Earabino C., Lipford J., Lindblom A., Tannergård P., Bollag R., Godwin A., Ward D., Nordenskjöld M., Fishel R., Kolodner R., Liskay R. 1994. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* 368:258–61.

Chen C., Rappailles A., Duquenne L., Huvet M., Guilbaud G., Farinelli L., Audit B., Aubenton-carafa Y., Arneodo A., Hyrien O., Thermes C. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. :447–457. DOI: 10.1101/gr.098947.109.

Chiappini F., Gross-Goupil M., Saffroy R., Azoulay D., Emile JF., Veillhan LA., Delvart V., Chevalier S., Bismuth H., Debuire B., Lemoine A. 2004. Microsatellite instability mutator phenotype in hepatocellular carcinoma in non-alcoholic and non-virally infected normal livers. *Carcinogenesis* 25:541–547. DOI: 10.1093/carcin/bgh035.

Cunningham JM., Christensen ER., Tester DJ., Kim CY., Roche PC., Burgart LJ., Thibodeau SN. 1998. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite

instability. *Cancer Research* 58:3455–3460.

Fishel R., Lescoe MK., Rao MRS., Copeland NG., Jenkins NA., Garber J., Kane M., Kolodner R. 1993. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 75:1027–1038. DOI: 10.1016/0092-8674(93)90546-3.

Gargiulo S., Torrini M., Ollila S., Nasti S., Pastorino L., Cusano R., Bonelli L., Battistuzzi L., Mastracci L., Bruno W., Savarino V., Sciallero S., Borgonovo G., Nystrom M., Bianchi-Scarr G., Mareni C., Ghiorzo P. 2009. Germline MLH1 and MSH2 mutations in Italian pancreatic cancer patients with suspected Lynch syndrome. *Familial Cancer* 8:547–553. DOI: 10.1007/s10689-009-9285-1.

Hansen RS., Thomas S., Sandstrom R., Canfield TK., Thurman RE., Weaver M., Dorschner MO., Gartler SM., Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* 107:139–44. DOI: 10.1073/pnas.0912402107.

Koren A., Polak P., Nemesh J., Michaelson JJ., Sebat J., Sunyaev SR., McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics* 91:1033–1040. DOI: 10.1016/j.ajhg.2012.10.018.

Lang GI., Murray AW. 2011. Mutation rates across budding yeast chromosome VI Are correlated with replication timing. *Genome Biology and Evolution* 3:799–811. DOI: 10.1093/gbe/evr054.

Miyakura Y., Sugano K., Konishi F., Ichikawa A., Maekawa M., Shitoh K., Igarashi S., Kotake K., Koyama Y., Nagai H. 2001. Extensive methylation of hMLH1 promoter region predominates in proximal colon cancer with microsatellite instability. *Gastroenterology* 121:1300–1309. DOI: 10.1053/gast.2001.29616.

Morganella, S. *et al.*, 2016. The topography of mutational processes in breast cancer genomes. Nature Communications, 7(May), p.11383. Available at: http://www.nature.com/doifinder/10.1038/ncomms11383.

Pauklin S., Sernández I V., Bachmann G., Ramiro AR., Petersen-Mahrt SK. 2009. Estrogen directly activates AID transcription and function. *The Journal of experimental medicine* 206:99–111. DOI: 10.1084/jem.20080521.

Stamatoyannopoulos JA., Adzhubei I., Thurman RE., Kryukov G V., Mirkin SM., Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nature genetics* 41:393–5. DOI: 10.1038/ng.363.

Supek F., Lehner B. 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* 521:81–84. DOI: 10.1038/nature14173.

Umar A., Boland CR., Terdiman JP., Syngal S., de la Chapelle A., Rüschoff J., Fishel R., Lindor NM., Burgart LJ., Hamelin R., Hamilton SR., Hiatt RA., Jass J., Lindblom A., Lynch HT., Peltomaki P., Ramsey SD., Rodriguez-Bigas MA., Vasen HFA., Hawk ET., Barrett JC., Freedman AN., Srivastava S. 2004. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute* 96:261–8. DOI: 10.1093/jnci/djh034.